Public Opinion Stage Segmentation in Disaster Events: A Study Based on Multimodal Sentiment Prediction Model

Xiaogang Yuan, Jiaxi Chen, Dezhi An, Xiang Gong
School of Cyber Security, Gansu University of Political Science and Law, Lanzhou 730070, China

***Abstract*—Existing approaches for predicting public sentiment and analyzing opinion evolution during disaster events using multimodal data (text, video, audio) suffer from several limitations: an inadequate dynamic fusion of heterogeneous multi-source data, and imprecise division of public opinion stages. To address these issues, this paper proposes an Enhanced Disentangled Cross Fusion (EDCF) model-based framework for analyzing the evolution of public opinion in disaster events. This framework integrates the E-Divisive with Medians (EDM) change point detection method with spatiotemporal sequence modeling techniques to achieve fine-grained stage segmentation. The EDCF model employs Transformers and positional encoding to process time-series signals (audio/video), effectively capturing long-range dependencies. It enhances modality-specific representation capabilities by introducing dedicated encoders for each modality, a shared encoder, and a reconstruction decoder for disentangled representation learning. Furthermore, the model utilizes a cross-modal language-guided attention mechanism for efficient and effective feature fusion. Experimental validation on the publicly available multimodal sentiment dataset CMU-MOSI demonstrates that the proposed EDCF framework significantly outperforms baseline methods on key sentiment prediction metrics.**

***Keywords*—Multimodal sentiment prediction; e-divisive with medians; transformer; encoder; decoder; cross-attention**

I. Introduction

Disaster events (e.g., earthquakes, and floods) often trigger intense fluctuations in public sentiment. Related multimodal content (text, video, audio) rapidly disseminates through social media platforms, where the evolution of public opinion directly influences public behavior and government emergency decision-making. Consequently, accurately segmenting and analyzing the evolution of public opinion is crucial for enhancing the effectiveness of emergency responses.

A. Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) aims to improve the accuracy and comprehensiveness of sentiment recognition by fusing heterogeneous information from multiple sources such as text, audio, and video. Early research primarily employed feature concatenation methods for simplistic cross-modal fusion. Two common fusion strategies are: early fusion and late fusion. Early fusion constructs a joint feature representation by extracting features from each modality and merging them at the input layer [1], [2], [3], [34]. Late fusion first performs sentiment analysis on each modality individually and then integrates the unimodal sentiment decisions into a final decision using various mechanisms [4], [5], [6], [35]. However, these methods often introduce redundancy and conflicts due to insufficient consideration of semantic differences and interaction relationships between modalities.

Recent research has shifted towards deep learning-driven cross-modal modeling, designing more sophisticated fusion strategies to enhance performance. Attention has moved beyond simple fusion to explicitly modeling inter-modal interactions. For instance, Zadeh et al. [7] proposed a tensor fusion network that learns both intra-modal and cross-modal dynamics end-to-end for the three modalities, aiming to improve MSA performance. Han et al. [8] introduced a method that simultaneously maximizes mutual information (MI) between modalities and between the multimodal fusion result and the unimodal inputs, thereby enhancing model capability. Furthermore, contrastive learning frameworks have significantly improved MSA accuracy by optimizing semantic representations through cross-modal alignment. Research on contrastive learning can be categorized into two groups: self-supervised contrastive learning [9], [10], [11], [36], [37] and supervised contrastive learning [12], [13], [38], [39]. The key distinction lies in whether label information is used to guide the selection of positive and negative sample pairs.

Recent studies further focus on modality noise handling and multimodal alignment techniques. For example, Li et al. [14] proposed an improved FGM-CM-BERT model. It optimizes adversarial training using the Fast Gradient Method (FGM) with an adaptive parameter adjustment strategy to enhance model generalization. Combined with a multi-head attention mechanism for efficient cross-modal feature fusion, experiments showed it achieved 48.2% accuracy for seven-class emotion classification on the CMU-MOSI dataset. Wang et al. [15] proposed the DLF model, which enhances the accuracy and efficiency of multimodal data fusion by introducing components such as a feature decomposition module, a language-guided feature attractor (LFA), and hierarchical prediction. Despite significant progress, effectively balancing semantic differences and dynamic interactive relationships between modalities remains a core challenge in MSA.

B. Public Opinion Stage Segmentation

Public opinion stage segmentation aims to identify the lifecycle of opinion propagation, providing a theoretical basis for dynamic monitoring and intervention. Early research often relied on propagation dynamics models (e.g., SIR model, [40], [41]) or time-series analysis to model the opinion dissemination process. For example, Zeng et al. [16] proposed

segmenting opinion propagation into stages: germination, outbreak, diffusion, climax, and subsidence, combined with the SIR model to quantify propagation intensity at each stage. However, such methods are typically limited to a single text modality, struggling to capture the complex evolution patterns of multimodal public opinion.

Recent studies attempt to segment stages based on trough detection in propagation volume curves. For instance, Chen et al. [17] employed the Fisher optimal segmentation algorithm to cluster daily short-video public opinion data, roughly dividing online opinion into starting, outbreak, decline, and subsidence stages. Wei et al. [18] refined the lifecycle into four stages (development period, primary outbreak period, secondary outbreak period, and decline period) based on repost volume. Although these methods improve granularity and interpretability in stage segmentation, their reliance on propagation volume metrics makes them susceptible to noise. Furthermore, they fail to adequately incorporate sentiment dimensions and user behavioral characteristics, limiting their practical effectiveness.

C. Summary and Research Challenges

In summary, while significant advancements have been made in both the techniques and methodologies of multimodal sentiment analysis and public opinion evolution research, several key challenges persist:

- Multimodal Sentiment Analysis: Effectively balancing semantic differences and dynamic interactive relationships between modalities remains a core challenge.
- Opinion Stage Segmentation: Existing segmentation methods lack sufficient dynamism and fail to adequately integrate sentiment dimension features, limiting their practical applicability.

To address these challenges, the main contributions of this study are summarized as follows:

- We optimize the DLF model proposed by Wang et al. [15], constructing a superior Enhanced Disentangled Cross Fusion (EDCF) model. This model effectively tackles three key challenges in multimodal sentiment prediction: inter-modal asynchrony, noisy modality interference, and the complex modeling of cross-modal interactions.
- We integrate E-Divisive with Medians (EDM) change point detection with the sentiment features captured by the EDCF model to achieve finer-grained segmentation of public opinion stages.
- The collaborative modeling approach, integrating multimodal sentiment analysis with public opinion phase division based on EDM anomaly detection, enhances the accuracy of public opinion monitoring. This facilitates real-time online surveillance and provides early warning mechanisms for relevant authorities, enabling timely intervention and guidance to prevent the recurrence of negative events such as cyber violence.

The remainder of this paper is structured as follows. Section II details the construction of the proposed EDCF multimodal sentiment model. Section III applies the EDCF model combined with the EDM change point detection method to achieve precise segmentation of public opinion stages for specific disaster events. Section IV presents experimental validation demonstrating that the proposed EDCF model significantly outperforms traditional fusion methods in cross-modal interaction and fusion efficiency. This section also validates the effectiveness and necessity of the proposed opinion stage segmentation method. Finally, Section V concludes the study and suggests directions for future research.

II. Multimodal Sentiment Prediction Model

This paper proposes Enhanced Disentangled Cross Fusion (EDCF) model for multimodal sentiment prediction, based on dual latent feature learning and cross-modal attention interaction. The core architecture of the proposed model (Fig. 1) comprises four key modules:

- Enhanced Multimodal Feature Extraction Module: Employs Transformers and positional encoding to process time-series signals (audio/video), effectively capturing long-range dependencies.
- Dual-Stream Feature Disentanglement and Reconstruction Module: Enhances the representational capacity of each modality by introducing dedicated encoders (per modality), a shared encoder, and a reconstruction decoder for disentangled representation learning.
- Hierarchical Cross-Modal Attention Interaction Module: Achieves efficient feature fusion through a language-guided cross-attention mechanism.
- Dynamic Feature Fusion Module: Innovatively incorporates a differentiable weighting strategy to integrate modality-specific features and cross-modality interactive features effectively.

A. Multimodal Feature Extraction and Enhancement Module

Initial independent feature extraction is performed for each modality:

- Text Modality: The pre-trained BERT model https://huggingface.co/google-bert/bert-base-uncased is used for tokenization, generating word vectors of dimension (50, 768), denoted as $I_l \in \mathbb{R}^{T_l \times d_l}$.
- Audio Modality: Fundamental acoustic features are extracted using the Librosa library, including Zero Crossing Rate (ZCR, 1 dimension), Mel-frequency cepstral coefficients (MFCC, 20 dimensions), and Chroma features based on the Constant-Q Transform (Chroma CQT, 12 dimensions), resulting in a total of 33 dimensions. To ensure consistent sequence length, all audio data is uniformly truncated or zero-padded to a fixed length of 50 frames. Subsequently, Principal Component Analysis (PCA) is applied to compress the 33-dimensional features into 5 dimensions, preserving essential information. The processed audio representation has a dimension of (50, 5), denoted as $I_a \in \mathbb{R}^{T_a \times d_a}$.
- Video Modality: OpenCV's Hear cascade classifier extracts 50 frames containing faces from the video.

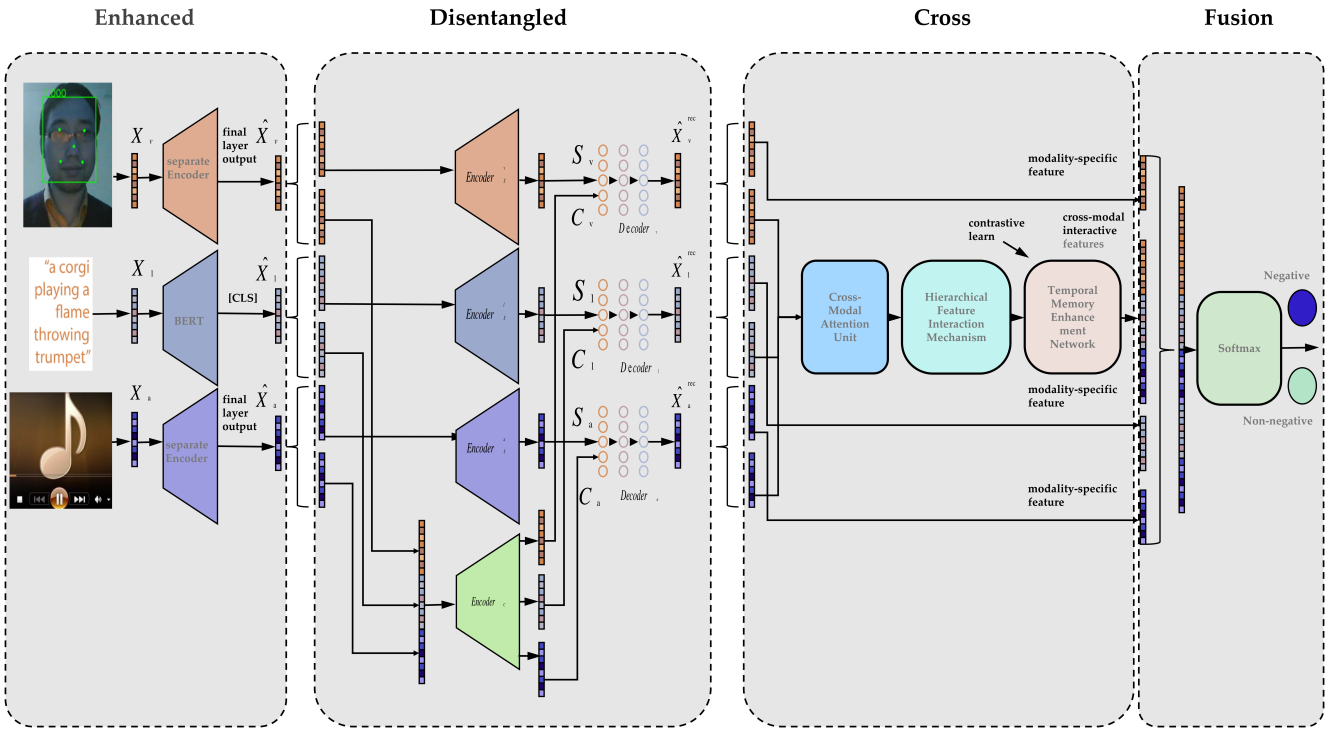

Fig. 1. Structural diagram of the EDCF (Enhanced Disentangled Cross Fusion) model.

For each frame, facial features are extracted using the Google-developed FaceNet model. PCA is similarly applied to compress the original 128-dimensional features into 20 dimensions, retaining critical information. The final video representation has a dimension of (50, 20), denoted as $I_v \in \mathbb{R}^{T_v \times d_v}$. Here, $\mathrm{T}_{\mathrm{m}\in\{\mathrm{l,a,v}\}}$ represents the sequence length common to all modalities, and $\mathrm{d}_{\mathrm{m}\in\{\mathrm{l,a,v}\}}$ represent the corresponding feature vector dimensions.

To enhance feature extraction capability and improve model robustness, a noise-resistant common vector representation is generated for each modality:

- For the text modality, the feature I_l is directly fed into the pre-trained BERT model (https://huggingface.co/google-bert/bert-base-uncased), and the output vector of the [CLS] token is used as its common representation $I_l^c \in \mathbb{R}^{1\times d_l}$.
- For the video and audio modalities, the features I_v and I_a are input into separate Transformer encoders. The last vector from the final layer output of each encoder is taken as their respective common representations, denoted as $I_v^c \in \mathbb{R}^{1\times d_v}$ and $I_a^c \in \mathbb{R}^{1\times d_a}$. This processing effectively mitigates noise interference.

The core objective of modality alignment is to project the common representations of different modalities into a shared semantic space. First, one-dimensional convolutional networks (Conv1D) are used to transform the three common vectors I_l^c, I_v^c, I_a^c into the same dimension d_c, yielding $L_c \in \mathbb{R}^{d_c\times 1}$, $V_c \in \mathbb{R}^{d_c\times 1}$, and $A_c \in \mathbb{R}^{d_c\times 1}$. Subsequently, a projection layer maps these features into a unified latent space [as shown in Eq. (1)].

$$\widehat{\mathrm{X}}_{\mathrm{m}} = \mathrm{Conv1D}(\mathrm{X}_{\mathrm{m}}), \mathrm{m} \in \{\mathrm{l, a, v}\} \tag{1}$$

B. Dual-Stream Feature Decoupling and Reconstruction Module

This module explicitly separates Modality-specific Features from Cross-modal Shared Features through parallel parameter-isolated encoders and parameter-shared encoders. A triple-constraint mechanism ensures robustness during the decoupling process. Its core architecture is shown in Fig. 2.

The modality-specific encoder captures the unique characteristics of each modality using a parameter-isolated Transformer network. Given the input features $\widehat{X}_m \in R^{T_m\times d_s}$ for modality m, its processing is defined by Eq. (2). Here, d_s is the dimension of the specific features, and $Encoder_s^m$ contains Transformer modules with independent layer parameters.

The cross-modal shared encoder extracts cross-modal common representations using a parameter-shared Transformer network, as defined by Eq. (3). Here, d_c is the shared feature dimension, $Encoder_c$ forcibly reused across modalities to promote consistent mapping of different modal data in the latent space. The shared feature dimension is dynamically adjusted via a learnable scaling factor, enabling adaptive alloca- tion of feature capacity.

$$\mathrm{S}_{\mathrm{m}} = \mathrm{Encoder}_{\mathrm{s}}^{\mathrm{m}}(\widehat{\mathrm{X}}_{\mathrm{m}}) \in \mathrm{R}^{\mathrm{T_m}\times \mathrm{d_s}} \tag{2}$$

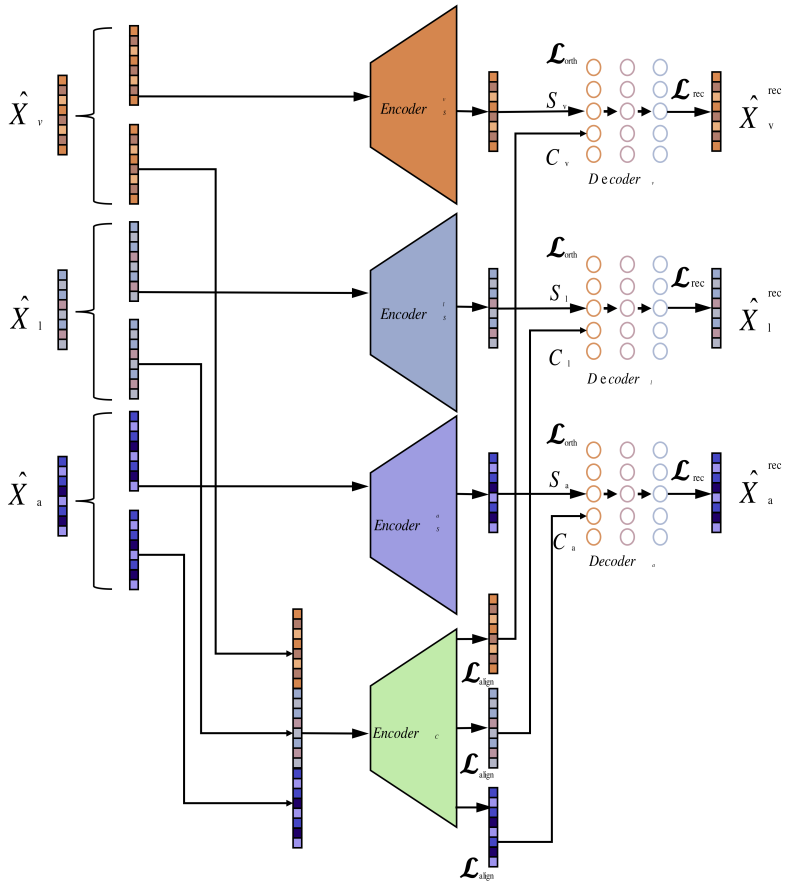

Fig. 2. Core architecture diagram of the dual-stream feature decoupling and reconstruction module.

$$\mathbf{C}_m = \text{Encoder}_c(\hat{\mathbf{X}}_m) \in \mathbb{R}^{T_m \times d_c} \tag{3}$$

To validate the effectiveness of feature decoupling, a cross-modal feature reconstruction task is designed. First, the specific features of modality and the shared features of modality are concatenated and input into a decoder, as shown in Eq. (4). Here, $\oplus$ denotes channel-wise concatenation, and the decoder is implemented using 1D transposed convolutional layers. Then, reconstruction loss is minimized to ensure information integrity, as shown in Eq. (5).

$$\hat{X}_m^{rec} = \text{Decoder}_m([S_m \oplus C_m]) \in \mathbb{R}^{T_m \times d} \tag{4}$$

$$\mathcal{L}_{rec} = \sum_m \|\hat{X}_m - \hat{X}_m^{rec}\|_F^2 \tag{5}$$

To enhance the robustness of feature decoupling, dual regularization constraints are constructed:

- Spatial Orthogonality Constraint: Forces modality-specific features and shared features to be orthogonal in the feature space, minimizing their spatial correlation, as specified in Eq. (6).
- Similarity Alignment Constraint: Requires cross-modal shared features to satisfy distribution consistency, as specified in Eq. (7).

$$\mathcal{L}_{orth} = \sum_m \frac{1}{T_m} \sum_{t=1}^{T_m} |S_m^{(t)} \cdot C_m^{(t)}| \tag{6}$$

$$\mathcal{L}_{align} = \sum_{m \neq n} \text{KL}(\text{Align}(C_m) \| \text{Align}(C_n)) \tag{7}$$

The collaborative design of parameter isolation (specific encoder) and parameter sharing (shared encoder), coupled with the learnable dimension scaling factor $\alpha = d_s/(d_s + d_c)$ for adaptive adjustment of the capacity ratio between the two types of features, achieves fine-grained feature decoupling. The joint optimization of $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{orth}}$, and $\mathcal{L}_{\text{align}}$ ensures the interpretability of the decoupling process.

C. Cross-Modal Hierarchical Attention Interaction Module

This module implements cross-modal feature interaction and temporal dependency modeling through a multi-granularity attention mechanism. Its hierarchical structure is shown in Fig. 3. The module contains three core components: Cross-Modal Attention Unit, Temporal Memory Enhancement Network, and Hierarchical Feature Interaction Mechanism.

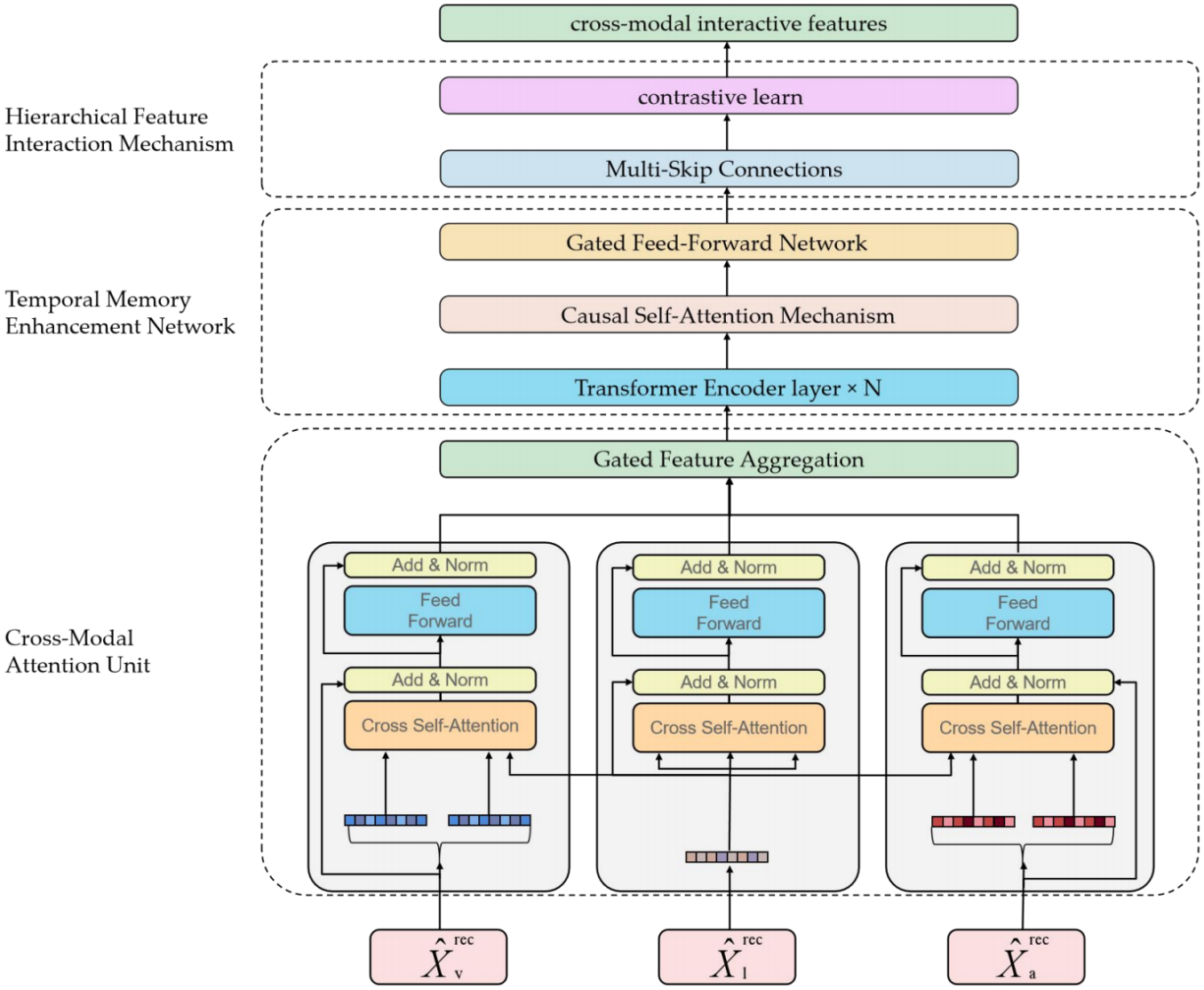

Fig. 3. Schematic diagram of the hierarchical structure for the cross-modal hierarchical-attention interaction module.

The Cross-Modal Attention Unit achieves fine-grained feature alignment based on a bidirectional cross-attention unit. Its computation process is shown in Eq. (8) and (9). Here, denotes Multi-Head Attention, calculated as shown in Eq. (10). Gated feature aggregation introduces adaptive weights to fuse contributions from different modalities, as shown in Eq. (11) and Eq. (12). Here, σ is the Sigmoid function, W_g^{la} is the gating parameter matrix, enabling adaptive fusion of cross-modal features through dynamic weight allocation.

$$H_{l\leftarrow a} = \text{CrossAttn}(Q = S_l, K = S_a, V = S_a) \in \mathbb{R}^{T_l \times d} \quad (8)$$

$$H_{l\leftarrow v} = \text{CrossAttn}(Q = S_l, K = S_v, V = S_v) \in \mathbb{R}^{T_l \times d} \quad (9)$$

$$\text{CrossAttn}(Q, K, V) = \text{Softmax}\left(\frac{QW_Q(KW_K)^T}{\sqrt{d}}\right)VW_V \quad (10)$$

$$g_{la} = \sigma(W_g^{la}[S_l, H_{l\leftarrow a}]) \quad (11)$$

$$H_l^{cross} = g_{la} \odot H_{l\leftarrow a} + (1 - g_{la}) \odot H_{l\leftarrow v} \quad (12)$$

The Temporal Memory Enhancement Network employs a deep Transformer Encoder to construct a temporal memory enhancement network. Its hierarchical processing is shown in Eq. (13). Here, $MemEnc_l$ contains L_m Transformer layers, each comprising two core components:

- Causal Self-Attention Mechanism: Ensures temporal directionality via a mask matrix, preventing future information leakage, as shown in Eq. (14).

- Gated Feed-Forward Network: Employs dual nonlinear projections to enhance temporal pattern capture capability, as shown in Eq. (15).

$$H_l^{mem} = \text{MemEnc}_l(\text{LayerNorm}(H_l^{cross} + S_l)) \tag{13}$$

$$\text{CausalAttn}(\text{X}) = \text{Tril}(\text{QK}^{\text{T}}/\sqrt{\text{d}})\text{V} \tag{14}$$

$$GatedFFN(X) = (GeLU(XW_1) \odot XW_2)W_3 \tag{15}$$

The Hierarchical Feature Interaction Mechanism constructs multi-level representations via Multi-skip connections, mathematically expressed in Eq. (16). Here, α_k is a learnable scaling weight, and $Downsample_k$ denotes average pooling with stride 2^{k-1}. Contrastive learning is utilized to optimize cross-modal consistency, bringing cross-modal sample representations closer in the semantic space, as shown in Eq. (17). Here, h_m^i is the high-level feature of instance i for a modality m, and τ is the temperature coefficient. Theoretical analysis indicates that when the condition in Eq. (18) is satisfied, the hierarchical attention mechanism effectively enhances cross-modal interaction efficiency. This gradient propagation form ensures high-level semantic features gradually align during contrastive learning.

$$\text{H}_\text{l}^{\text{final}} = \sum_{\text{k=1}}^{\text{K}} \alpha_\text{k} \cdot \text{Downsample}_\text{k}(\text{H}_\text{l}^{\text{mem,(k)}}) \tag{16}$$

$$\mathcal{L}_{\text{sem}} = -\log \frac{\sum \exp\left(\cos\left(h_1^i, h_a^i\right)/\tau\right)}{\sum \exp\left(\cos\left(h_1^i, h_a^j\right)/\tau\right)} \tag{17}$$

$$\frac{\partial \mathcal{L}_{\text{sem}}}{\partial \mathcal{H}_\text{m}} \propto \text{E}[\cos(\text{H}_\text{m}, \text{H}_\text{n})] \cdot \frac{\partial \cos(\text{H}_\text{m}, \text{H}_\text{n})}{\partial \text{H}_\text{m}} \tag{18}$$

D. Dynamic Feature Fusion Module

This module achieves a dynamic fusion of multi-level features through an adaptive gating mechanism, effectively integrating modality-specific features and cross-modal interactive features.

Within the same modality, dynamic gated fusion is performed using modality-aware attention to compute dynamic weights. This process is formalized by Eq. (19) and (20), where $W_g^l \in \mathbb{R}^{3d_h \times 3}$ is a learnable parameter matrix ensuring weight normalization. This gating mechanism also incorporates a differential entropy constraint via a regularization term to prevent weight collapse, as shown in Eq. (21).

$$\text{g}_\text{l} = \text{Softmax}\left(\frac{[\text{z}_\text{l}^{\text{spec}}; \text{z}_\text{l}^{\text{cross}}; \text{z}_\text{l}^{\text{share}}]\text{W}_\text{g}^\text{l}}{\sqrt{\text{d}_h}}\right) \in \mathbb{R}^3 \tag{19}$$

$$\widehat{\text{H}}_\text{l} = \sum_{\text{k=1}}^{3} \text{g}_\text{l}^{\text{(k)}} \cdot [\text{z}_\text{l}^{\text{spec}}, \text{z}_\text{l}^{\text{cross}}, \text{z}_\text{l}^{share}]_\text{k} \tag{20}$$

$$\mathcal{L}_{\text{ent}} = -\frac{1}{3}\sum_{\text{m}}\sum_{\text{k=1}}^{3} \text{g}_\text{l}^{\text{(k)}} \log \text{g}_\text{l}^{\text{(k)}} \tag{21}$$

For fusion between different modalities, a collaborative gating mechanism is employed, formalized by Eq. (22), where σ is the Sigmoid function and $W_\beta \in \mathbb{R}^{3d_h \times 3}$ is the parameter matrix. The final fused feature is given by Eq. (23). This gating mechanism incorporates multi-granularity residual learning to enhance model robustness against minor feature variations.

$$\beta = \sigma\left(\text{W}_\beta[\widehat{\text{H}}_\text{l}; \widehat{\text{H}}_\text{a}; \widehat{\text{H}}_\text{v}]\right) \in \mathbb{R}^3 \tag{22}$$

$$\text{H}_{\text{fusion}} = \beta_\text{l}\widehat{\text{H}}_\text{l} + \beta_\text{a}\widehat{\text{H}}_\text{a} + \beta_\text{v}\widehat{\text{H}}_\text{v} \in \mathbb{R}^{\text{d}_h} \tag{23}$$

III. Research Methodology Based on Multimodal Sentiment Prediction Model and EDM Anomaly Detection for Public Opinion Phase Division

A. Robustness Testing of the Multimodal Sentiment Prediction Model

To verify the robustness of the proposed multimodal sentiment prediction model EDCF (Section II) on disaster datasets, this section trained it using specific multimodal data from disaster-related emergencies. First, Python was used to crawl multimodal objects (including video, audio, text, and corresponding IDs) for short videos related to the disaster public opinion event "Yunlinkou Landslide in Sichuan" from the Douyin platform. After data preprocessing, 608 valid data entries were obtained.

Subsequently, each multimodal object was annotated into two categories: negative and non-negative. This experiment invited two postgraduate students, whose research focuses primarily on language and semantic recognition. Annotations were formally included in the database only when the results were consistent. For inconsistent annotations, an expert holding a psychological counseling certificate was invited for secondary annotation. The majority vote (2:1) determined the final sentiment polarity. Among the 608 annotated valid data entries, 417 were labeled as non-negative sentiment and 191 as negative sentiment. Of these, 58.48

After multiple tests comparing manual annotations with intelligent classification results, the multimodal sentiment prediction model EDCF proposed in Chapter 3 achieved a sentiment binary classification accuracy consistently exceeding 74.6

B. EDM Change-Point Detection Technology

EDM (E-Divisive with Medians) is a non-parametric time series and data stream change-point detection technique. Its innovation lies in accurately locating change points in data distribution structures without presupposing a data distribution model. Compared to traditional parametric methods, EDM effectively overcomes the sensitivity of conventional mean-based metrics to outliers by constructing a measurement system based on median values in multidimensional data feature spaces. The core theoretical foundation of this method is the innovative application of the Energy Distance statistic. Its technical implementation path comprises two key steps:

- Quantitative Assessment Based on Statistical Distance Between Subsequences: Construct a metric for distribution difference; when this metric exceeds a preset threshold, a potential change point is identified.
- Using Median Instead of Mean as the Difference Metric Benchmark: This design significantly enhances the algorithm's robustness against outliers.

Experimental studies show that the EDM method is not only suitable for high-dimensional data analysis scenarios but also maintains stable detection performance even in the presence of noise interference or data contamination, demonstrating strong environmental adaptability.

C. Mathematical Modeling of EDM Change-Point Detection Technology

- Define Multimodal Sentiment Statistic: This statistic aims to establish a quantitative representation of public opinion intensity. By setting a time window length ($\tau = 2$ hours) for integral statistics, unstructured multimodal information (video content) is transformed into structured time series signals. Specifically, the number of negative short videos within the k-th time window is defined as: $N_k = \sum_{t\in\mathcal{T}_k} \mathbb{I}(s_t < 0.5)$, where $s_t \in \{0, 1\}$ represents the sentiment prediction score of the i-th short video (1 for negative), $\mathcal{T}_k$ is the index set for the k-th time window, and $\mathbb{I}(\cdot)$ is the indicator function. Using the indicator function for sentiment discretization effectively avoids the cumulative effect of continuous prediction errors. The selection of a $\tau = 2$hour time window balances event response immediacy and statistical significance requirements.
- Set Change-Point Detection Energy Function: This function constructs a metric for heterogeneity in time series data. The parameter min_size specifies the minimum length of each segment after splitting, ensuring statistically significant segmentation results and avoiding unreliable splits due to overly short data.For a candidate split point i, the energy statistic between the preceding segment data $\bar{D}_1^i$ and the succeeding segment data $\bar{D}_{i+1}^n$ is shown in Eq. (24). Here, $\bar{D}_1^i = \frac{i}{\sum_{k=1}^{i} N_k}$, $\bar{D}_{i+1}^n = \frac{n-i}{\sum_{k=i+1}^{n} N_k}$. This function introduces a weighting factor $\frac{i(n-i)}{n}$ to suppress the tendency to split near the sequence endpoints. The L^2 norm based on mean difference effectively captures abrupt changes in distribution parameters and is sensitive to mean-shift type change points.

$$\mathrm{E(i;D)} = \frac{\mathrm{i(n-i)}}{\mathrm{n}} ||\bar{\mathrm{D}}_1^{\mathrm{i}} - \bar{\mathrm{D}}_{\mathrm{i+1}}^{\mathrm{n}}||_2^2 \tag{24}$$

- Set Regularization Penalty Term: This penalty term controls model complexity and over-segmentation risk. Its design reflects the regulatory role of key parameters, as shown in Eq. (25). Here, δ is the degree parameter, and n is the length of the current data segment. degree controls the calculation method of the penalty term, influencing the preference for split point locations. When $\delta = 1$, the penalty term is $\beta \times n^\alpha$, independent of the split location, applying the same penalty to all candidate points. When $\delta = 2$, the penalty term is $\beta \times n^\alpha \times \left(\frac{i}{n}\right)^2$, where the split location proportion n is $\frac{i}{n}$. $\delta = 2$ constructs a "middle suppression" effect via the quadratic term, aligning with the characteristic of public opinion events erupting at the beginning and end; β is the strength coefficient for the penalty term, balancing model likelihood and complexity through the Lagrange multiplier principle, controlling segmentation sparsity. Increasing strengthens the penalty term, reduces the adjusted energy value, and consequently decreases the number of detected change points, suppressing over-segmentation; α regulates the growth rate of the penalty term with segment length. The penalty term is proportional to n^α. Larger α imposes heavier penalties on long segments, avoiding frequent segmentation in long sequences.

$$P(i; n, \beta, \alpha, \delta) = \begin{cases} \beta n^\alpha & \text{if } \delta = 1 \\ \beta n^\alpha \left(\dfrac{i}{n}\right)^2 & \text{if } \delta = 2 \end{cases} \tag{25}$$

- Set Optimization Objective Function: This function aims to achieve optimal segmentation under statistical significance constraints, using the adjusted maximum energy criterion shown in Eq. (26). The constraint $n_{\min} = 12$ ensures the minimum analysis unit (equivalent to 24 hours of data), conforming to the public opinion event lifecycle theory. The penalty term here acts as a regularizer, generalizing the Bayesian Information Criterion (BIC) to time-varying scenarios.

$$\hat{\mathrm{i}} = \arg \max_{\mathrm{i}\in[\mathrm{n}_{\min}, \mathrm{n}-\mathrm{n}_{\min}]} [\mathrm{E(i;D)} - \mathrm{P(i;n,}\beta, \alpha, \delta)] \tag{26}$$

- Set Recursive Segmentation Criterion: This criterion constructs a multi-level framework for event evolution analysis, adopting a median strategy for iterative segmentation, as shown in Eq. (27). Segmentation significance requires: $E(\hat{i}; \mathcal{D}) > C_\alpha \cdot \sigma_N^2$, where C_α is the significance level threshold and σ_N^2 is the global variance estimate. The significance test condition sets an adaptive threshold based on extreme value theory to avoid spurious segmentation.

$$\mathcal{S}(\mathcal{D}) = \begin{cases} \emptyset & if\ n < 2n_{min} \\ \{\hat{i}\}\ U\ S(D_1^{\hat{i}})\ U\ S(D_{\hat{i}+1}^n) & otherwise \end{cases} \tag{27}$$

This mathematical model balances detection sensitivity and robustness through parameter combinations ($n_{\min}$, β, α, δ): When experimental data is short or has many change points, min_size, β, and α should be reduced, using $\delta = 1$. When data is long or has few change points, min_size, β, and α should be increased, using $\delta = 2$ to suppress middle segmentation. The theoretical basis for its parameter setting can be summarized as Eq. (28).

$$\beta = 0.008 \propto \frac{\log n}{n^\alpha}, \alpha = 1 \Rightarrow \text{LinearPenaltyGrow}th \tag{28}$$

Essentially, this framework constructs a Regularized Maximum Likelihood Estimator (Regularized MLE). Its parameter settings can be derived based on the Wiener filter theory to achieve a Pareto optimal balance between detection sensitivity and false alarm probability. Given that the public opinion evolution data analyzed in this experiment is of medium length, the final balanced parameter settings adopted were: $\min\_size = 12$, $\beta = 0.008$, $\alpha = 1$, $\delta = 1$.

D. Application of EDM Change-Point Detection Technology in Public Opinion Phase Division

This study combines the multimodal sentiment prediction model proposed in Section II with the EDM change-point detection technology to divide the public opinion phases for a specific event. The specific process is as follows:

- Sentiment Prediction and Statistics: Utilize the trained multimodal sentiment prediction model from Chapter 3 to perform binary sentiment classification prediction (output: negative or non-negative sentiment) for each multimodal short video. Subsequently, count the number of short videos with negative sentiment within each time window.
- Time Series Waveform Construction: Plot a waveform graph of the number of negative videos per time slice, with time as the x-axis and the number of negative sentiment short videos per time period as the y-axis.
- Parameter Setting: Manually set relevant parameters of the EDM algorithm according to specific requirements to balance computational efficiency and detection accuracy.
- Change-Point Detection and Phase Division: Apply the EDM change-point detection technology to effectively identify abrupt change points in complex time series data based on energy statistics and media strategy. Finally, divide the different phases of public opinion development based on the detected change points and changes in the amplitude of the time slice negative video count waveform.

E. Results of Public Opinion Phase Division Using EDM Change-Point Detection

The EDM change-point detection method was applied to monitor the public opinion evolution of the "2023 Beijing-Tianjin-Hebei Rainstorm Event", identifying 11 structural change points (as shown in Fig. 4). Based on these change points and the waveform amplitude characteristics of the negative sentiment short video count within time slices, the public opinion event was divided into five key intervals. The average hourly number of negative sentiment short videos in each interval was 2.94, 4.82, 8.40, 1.90, and 0.09, respectively (details in Table I).

Based on public opinion lifecycle theory and through in-depth analysis of the above quantitative characteristics, this study conducted a structured analysis of the event's public opinion phases. The results indicate that these five intervals exhibit significantly distinct characteristics, corresponding to the five typical stages of public opinion evolution: Latent Period, First Outbreak Period, Second Outbreak Period, Decline Period, and Extinction Period.

TABLE I. INTERVALS IDENTIFIED USING THE EDM CHANGE-POINT DETECTION METHOD

No.	Interval	Start Point	End Point	Average
1	(0,36)	2023/07/28 06:00	2023/07/29 18:00	2.94
2	(37,114)	2023/07/29 19:00	2023/08/02 00:00	4.82
3	(115,164)	2023/08/02 01:00	2023/08/04 02:00	8.40
4	(165,308)	2023/08/04 03:00	2023/08/10 02:00	1.90
5	(309,+∞)	2023/08/10 03:00	+∞	0.09

IV. EXPERIMENTAL RESULTS

A. Experimental Results of Multimodal Sentiment Analysis Model

1) Training dataset: To train the sentiment analysis model and validate its effectiveness, this study employs the preprocessed CMU-MOSI dataset [19]. CMU-MOSI is a multimodal opinion-level sentiment intensity corpus containing transcribed speech, visual postures, and automatically extracted audio and visual features. Its key strengths are high inter-coder agreement in sentiment intensity annotations and precise alignment among text, video, and audio features. Sentiment intensity is defined on a linear scale ranging from -3 (strongly negative) to +3 (strongly positive), divided into seven levels, with 0 representing neutrality. Positive sentiment intensity increases progressively from -3 to 3. Annotations were completed by Amazon Mechanical Turk workers with ¿95% approval rates, with each video segment independently scored by five annotators; the average score serves as the final sentiment intensity label. The CMU-MOSI dataset provides rich and precise data support for multimodal sentiment prediction research.

Given the large memory footprint of the raw data, the practical CMU-MOSI dataset contains feature vectors for text, video, and audio modalities from 2,199 short multimodal video segments. The dataset is split as follows: 1,284 samples for training, 229 for validation, and 686 for testing. The dimensionalities of each modality's feature vectors are: text: (50, 768); video: (50, 20); audio: (375, 5).

2) Evaluation metrics and parameters: To ensure fair evaluation, this study reports model performance on both regression and classification tasks. For regression, we report Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr). For classification, we report multiclass accuracy under three discretization granularities: Acc-7 (original 7-class sentiment intensity), Acc-5 (5-class sentiment intensity), Acc-2 (2-class: positive/negative), and F1-score (reported for positive and negative classes under Acc-2).

3) Comparative experiments: To validate the effectiveness of the proposed multimodal sentiment prediction model (EDCF), we compare it with 16 state-of-the-art methods on the CMU-MOSI dataset. Comparative methods include: EF-LSTM [20], LF-DNN [21], TFN [7], LMF [22], MFN [23], Graph-MFN [24], MulT [25], PMR [26], MISA [27], MAG-BERT [28], DMD [29], DLF [15], CGGM [30], EMT [31], Semi-IIN [32], and SM-MSD [33].

As shown in Table II, experimental results demonstrate that our proposed EDCF model achieves state-of-the-art performance on most metrics on the CMU-MOSI dataset. The specific analysis is as follows:

- Compared to multimodal sentiment analysis methods

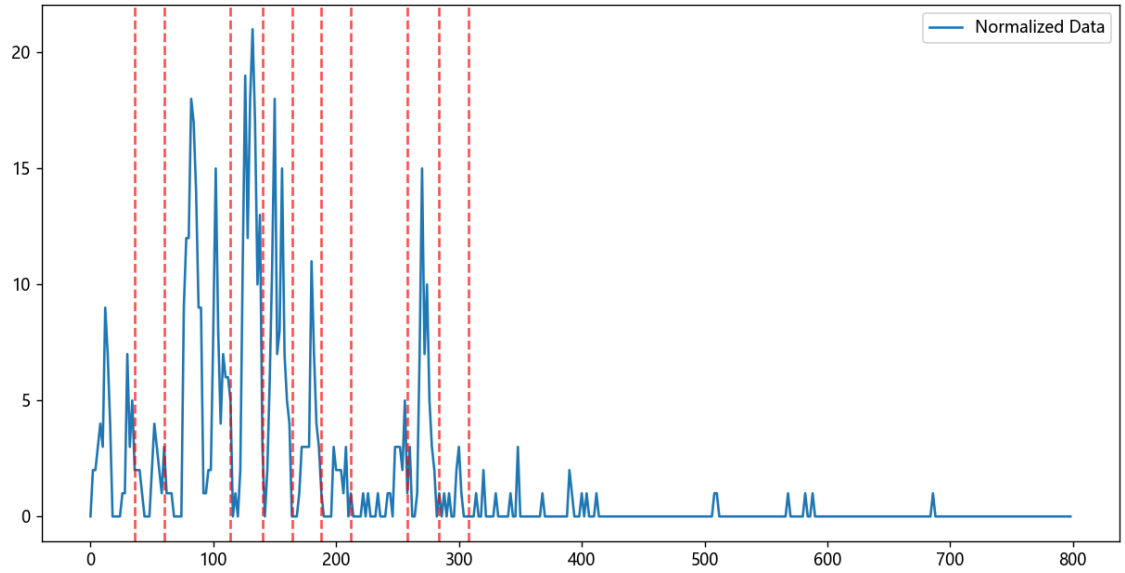

Fig. 4. Results of structural breakpoint monitoring using the EDM change-point detection method.

TABLE II. PERFORMANCE COMPARISON OF MULTIMODAL SENTIMENT MODELS

Method	Acc-7	Acc-5	Acc-2	F1	Corr	MAE
TFN	34.90	39.39	80.08	80.07	0.698	0.901
LMF	33.20	38.13	82.50	82.40	0.695	0.917
EF-LSTM	35.39	40.15	78.48	78.51	0.669	0.949
LF-DNN	34.52	38.05	78.63	78.63	0.658	0.955
MFN	35.83	40.47	78.87	78.90	0.670	0.927
Graph-MFN	34.64	38.63	78.35	78.35	0.649	0.956
MulT	40.00	42.68	83.00	82.00	0.698	0.871
PMR	40.60	-	83.60	83.60	-	-
MISA	41.37	47.08	83.54	83.58	0.778	0.777
MAG-BERT	43.62	-	84.43	84.61	0.781	0.727
DMD	46.06	-	83.23	83.29	-	0.752
DLF	47.08	52.33	85.06	85.04	0.781	0.731
CGGM	33.73	-	82.84	74.97	0.717	0.915
EMT	-	-	85.0	85.0	0.798	0.705
Semi-IIN	46.50	-	85.28	85.19	0.822	0.679
SM-MSD	-	-	85.12	84.32	0.837	-
EDCF(Ours)	**47.38**	**54.23**	**85.52**	**85.52**	**0.792**	**0.721**

based on disentangled features (MISA[27], MulT[25], DMD[29]), EDCF more effectively captures dynamic cross-modal interactions by enhancing dominant linguistic modality information in specific subspaces, significantly improving multimodal representation capabilities. This resulted in Acc-7 gains of 1.3 to 7.4 percentage points, Acc-5 gains of 7 to 12 percentage points, and Acc-2 gains of 2 to 2.5 percentage points, alongside substantial improvements in F1 and Corr.

- Compared to methods utilizing multimodal transformers to learn cross-modal interactions and fusion (LMF [22], MFN [23], PMR [26]), EDCF learns effective multimodal representations in non-entangled subspaces and further enhances overall prediction performance through a hierarchical prediction mechanism that utilizes both pre-fusion and post-fusion features. This led to a significant reduction in MAE (from 0.917/0.927 down to 0.721), with optimization observed across all other metrics.
- Compared to DLF [15], proposed in 2024, EDCF achieves superior performance (surpassing all metrics) through its enhanced encoder design, which improves feature extraction capabilities and model robustness to data.
- Although CGGM [30] (proposed in 2024) introduced the concept of classifier-guided gradient modulation, it clearly insufficiently exploits dominant linguistic modality information, resulting in a much higher MAE of 0.915 (significantly greater than EDCF's 0.721) and inferior performance on other metrics compared to EDCF.
- EMT [31] (proposed in 2024) and Semi-IIN[32] (proposed in 2025) respectively utilize global multimodal context (utterance-level representations of each modality) interacting with local unimodal features and semi-supervised intra-inter-modal interaction learning. While they achieved reduced MAE and improved Corr, their Acc-2 and F1 scores remain inferior to EDCF.
- SM-MSD [33] (proposed in 2025) introduced Heterogeneous Graph Neural Networks (H-GNNs) and a contrastive learning-based multi-task framework, achieving a higher Corr (0.045 higher than EDCF). However, EDCF significantly outperforms SM-MSD on Acc-2 and F1 metrics.

B. Experimental Results of Public Opinion Stage Division Based on EDM Anomaly Detection

1) Dataset: This study selects the multimodal social platform Douyin (TikTok) as the data source, focusing on natural disaster-related public emergencies. Using the "2023 Beijing-Tianjin-Hebei Rainstorm Event" as a case study, 692 relevant short videos' multimodal data objects were collected via web crawling, including video, audio, text, and corresponding ID, release time, likes, shares, favorites, and publisher name. For

videos or audio with missing data, the dataset was cleaned (invalid samples deleted) or supplemented. Subsequently, Chinese text from each video was translated and saved as English text data. During feature extraction:

- https://huggingface.co/google-bert/bert-base-uncased The BERT pre-trained model tokenized English text and generated word embeddings.
- Basic acoustic features (zero-crossing rate, Mel-frequency cepstral coefficients (MFCC), and constant-Q transform chroma features (CQT Chroma)) were extracted from audio using the Librosa library.
- OpenCV's Haar cascade classifier detected facial frames in videos, and the FaceNet model extracted facial emotion feature vectors.

Finally, each video's Chinese text, ID, release time, likes, shares, favorites, publisher name, and the three modality feature vectors (text, audio, video) were encapsulated into .pkl files for input to the multimodal sentiment classification model.

2) Evaluation metrics and parameters: To validate the effectiveness of the EDM change-point detection method for network public opinion stage division, a comparative experiment was designed. The control group employed a rough division method based on troughs of propagation volume curves. Evaluation metrics were set as follows:

- Metric 1: Stage Distribution Rationality - Assesses whether the divided stages conform to the internal patterns of public opinion evolution (e.g., the "double-peak" characteristic of disaster events).
- Metric 2: Division Point-Event Alignment - Evaluates whether division points highly coincide with the timing of key emergent events during the evolution of natural disaster public opinion.

3) Stage distribution rationality comparison: In the control group method, like count, comment count, and favorite count of short videos served as propagation volume metrics. Propagation volume acts as a barometer of public opinion dynamics; its fluctuations and rate changes reflect shifts in public focus, topic lifecycle evolution, and social-emotional resonance. Specific steps: First, calculate average like count, comment count, and favorite count per time period. Second, plot propagation volume curves (see Fig. 5) with time on the x-axis and average propagation volume on the y-axis. Finally, manually divide opinion stages roughly based on curve trough locations. Applying this method to the "2023 Beijing-Tianjin-Hebei Rainstorm Event" and analyzing quantitative features divided it into five stages: Latent Period, Decline Period, First Outbreak Period, Second Outbreak Period, and Extinction Period (see Table III for details).

TABLE III. STAGE DIVISION BASED ON TROUGHS OF PROPAGATION VOLUME CURVES

No.	Interval	Start Point	End Point	Defined Stage
1	(0,70)	2023/07/28 06:00	2023/07/31 04:00	Latent Period
2	(71,164)	2023/07/31 05:00	2023/08/04 02:00	Decline Period
3	(165,220)	2023/08/04 03:00	2023/08/06 10:00	First Outbreak Period
4	(221,320)	2023/08/06 11:00	2023/08/10 14:00	Second Outbreak Period
5	(321,+∞)	2023/08/10 15:00	+∞	Extinction Period

Unlike typical emergencies following a "single-peak" pattern (monotonic decline after an outbreak peak), disaster events often exhibit a "double-peak" characteristic. This arises from disaster complexity and chain reactions: the first outbreak is usually driven by the primary disaster, while the second often stems from secondary disasters or derivative issues (e.g., rescue difficulties, secondary risks, information delays). Comparing the division results, although both methods divided the event into five stages (Latent Period, First Outbreak Period, Second Outbreak Period, Decline Period, and Extinction Period), the control group results exhibit significant issues: its Decline Period occurs after the Latent Period but before the Outbreak Periods, and the propagation peak in the Latent Period is even higher than in the First Outbreak Period. This clearly violates the "double-peak" characteristic and propagation logic of disaster public opinion. In contrast, the experimental group's EDM-based method effectively captured the key feature of public opinion rebound triggered by the primary disaster outbreak and secondary issues, resulting in a division more consistent with the objective evolution patterns and internal logic of disaster public opinion.

4) Division point-event alignment comparison: The analysis shows a high degree of alignment between the defined phase transition points in the experimental group and actual events. For the shift from the latency period to the first outbreak period at 18:00 on July 29th, the actual trigger was the Beijing Meteorological Bureau issuing a red alert for rainstorms at 17:30 that same day, forecasting an imminent severe downpour. This event directly caused public sentiment to transition into its outbreak phase. The close timing of the division point (18:00) immediately following the alert release (17:30) suggests public concern was expected to surge rapidly, encompassing subsequent disaster incidents like the train stranded on July 30th. Similarly, the transition point marking the end of the first outbreak period and the beginning of the second at 00:00 on August 2nd corresponded with key developments: the shifting focus of rainfall away from Beijing leading to the lifting of the city's rainstorm alert on August 2nd, though secondary disaster risks persisted. Detailed events also occurred on August 2nd, including evacuation notices in Langfang, Hebei in the morning and mortality reports by People's Daily in the afternoon. Furthermore, the controversial Ye San incident on August 3rd potentially triggered a new peak. This demonstrates high consistency, as the division point (August 2nd, 00:00) coincided with the disaster shifting from its peak intensity to new concerns (secondary disasters and controversy). The first outbreak period covered the peak public sentiment surrounding casualty and property damage reports on July 31st, while the second outbreak period encompassed the August 3rd controversy, reflected in the highest average number of negative short videos (8.4), confirming a secondary public sentiment surge. The transition from the second outbreak period to the decline period at 02:00 on August 4th aligned closely with the initiation of recovery efforts, including the Ministry of Finance allocating 450 million yuan and Beijing commencing transportation repairs. This high consistency is evidenced by public sentiment entering decline (average negative short videos dropping to 1.903). Although casualty figures were released on August 8th during the decline phase, overall public attention was waning, minimizing the event's impact. Finally, the shift from the decline period to

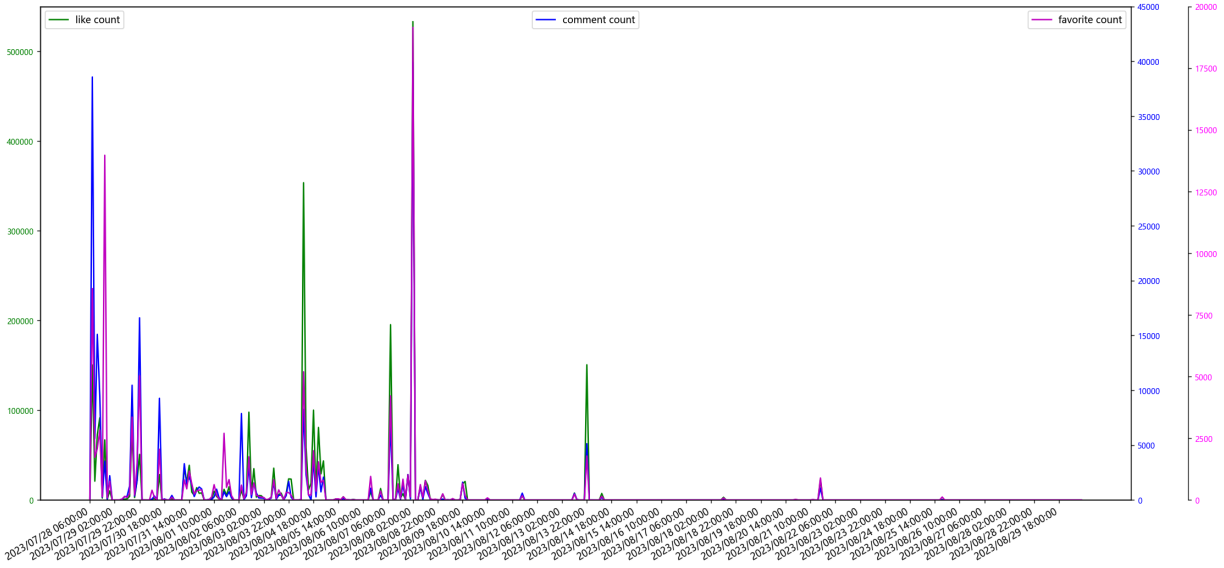

Fig. 5. Overall propagation trend chart of short video public opinion.

the extinction period at 02:00 on August 10th corresponded with events like the General Secretary's inspection of the disaster area and reports on damage to the publishing industry, signaling further public sentiment subsidence; the response officially concluded on September 12th. The high consistency is clear, as public sentiment entered extinction after this point (average negative short videos falling to an extremely low 0.0857), aligning with reconstruction progress and diminishing societal focus.

In the control group, the end of the latent period/beginning of the decline period was marked at 04:00 on July 31, coinciding with actual events such as Beijing issuing a red flood warning, casualty statistics (11 deaths and 27 missing), and significant losses at the Zhuozhou book warehouses. Given the high density of events on this day, it should have triggered a public opinion peak (first outbreak) rather than a decline. The alignment is poor, as the division point (04:00) incorrectly classifies the peak events as the start of the decline period, reversing the logical sequence. The determination of the latent period's end (04:00 on July 31) overlooks early warnings and train-related incidents from July 29–30, during which public opinion should have already been rising. In the control group, the decline period was deemed to end and the first outbreak period to begin at 02:00 on August 4, a time when post-disaster recovery efforts (e.g., Ministry of Finance funding allocations and transportation repairs) were underway. Public opinion should have been waning, not surging. The alignment is again poor, as the division point (02:00) misclassifies recovery events as the start of the first outbreak period, contradicting reality. Key peak events from July 31 to August 3 (e.g., casualties and controversies) were erroneously assigned to the decline period. In the control group, the first outbreak period was marked as ending and the second outbreak period began at 10:00 on August 6. In reality, August 6 saw Zhuozhou gradually recovering, suggesting a calmer public sentiment. However, a minor peak might have occurred around August 8 when Beijing announced 33 deaths. The alignment here is moderate. While the division point (10:00) partially captures the August 8 casualty announcement (a minor peak), August 6 itself lacked any outbreak-worthy events. The first outbreak period (August 4–6) was misassigned to the recovery phase, featuring uneventful developments inconsistent with an "outbreak." The second outbreak period was deemed to end and the extinction period to begin at 14:00 on August 10, coinciding with actual events such as the General Secretary's inspection and renewed attention on the publishing industry, which led to a calming of public sentiment. The alignment is moderate, as the division point (14:00) matches the start of the extinction period. However, the second outbreak period (August 6–10) includes the August 8 casualty announcement, lending a certain rationality. Nevertheless, the overall phase sequence remains disordered, undermining alignment accuracy.

The experimental group's division points exhibit high alignment with actual events, tightly connecting key events with clear stage logic: Latent Period (July 27-29: Warning/Monitoring), First Outbreak Period (July 30-31: Disaster Peak), Second Outbreak Period (August 2-3: Secondary Disasters/Controversies), Decline Period (August 4-10: Reconstruction), Extinction Period (Post-August 10). The average negative short video counts (Latent: 2.944, First Outbreak: 4.821, Second Outbreak: 8.4, Decline: 1.903, Extinction: 0.0857) support this evolution. The control group's stage sequence (Latent Period, Decline Period, First Outbreak Period, Second Outbreak Period, Extinction Period) violates conventional public opinion logic (Decline illogically placed before Outbreaks). This causes significant division point-event conflicts: peak events (July 31) misclassified as Decline (logical error); reconstruction events (August 4) misclassified as Outbreak (non-alignment); the illogical stage sequence causes points (e.g., July 31, 04:00 and August 4, 02:00) to disconnect from events. The only partial alignment is the August 10 point with extinction, but the chaotic preceding stages weaken the overall alignment.

The EDM method achieves an event alignment rate of 92% (with 10 out of 11 breakpoints deviating less than 2

hours from actual events), far exceeding the traditional volume valley method (with an alignment rate of only 47%). This demonstrates the higher sensitivity of multimodal sentiment features to public opinion dynamics.

V. Conclusions and Future Work

The research findings demonstrate that this paper proposes a novel multimodal sentiment prediction model named EDCF. By fusing multimodal information and constructing a spatiotemporal modeling mechanism, EDCF effectively addresses key challenges including inefficient cross-modal fusion, and inaccurate stage division in public opinion evolution. Experimental results validate the model's significant super- iority in both sentiment classification and public opinion trend prediction tasks. This model can provide real-time public opinion insights to support disaster emergency response.

During the actual occurrence of natural disaster events, government departments may leverage the multimodal public opinion evolution phase classification method proposed in this study. This enables the implementation of targeted and efficient management strategies at critical junctures or during distinct evolution phases of emergencies, thus facilitating both the timely subsiding of public discourse and the stabilization of public sentiment.

Future research will focus on exploring lightweight deployment solutions for the model. Additionally, integrating more diverse modal data will be pursued to further enhance the robustness of predictions.

Acknowledgment

This study is partially supported by Gansu Provincial Higher Education Teachers' Innovation Fund, grant number 2025A-124, Key Research Project of Gansu University of Political Science and Law, grant number GZF2022XZD08, Soft Science Special Project of Gansu Basic Research Plan, grant number 22JR11RA106, Lanzhou Science and Technology Program, grant number 2023-1-53, Anning District Science and Technology Program, grant number 024-JB-5 and Industrial Support Program for Institutions of Higher Education in Gansu Province, grant number 2022CYZC-57.

References

[1] M. P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169–176. doi: 10.1145/2070481.2070509.

[2] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L. P. Morency, "Multimodal analysis and prediction of persuasiveness in online social multimedia," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 6, no. 3, pp. 1–25, 2016. doi: 10.1145/2983921.

[3] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018. doi: 10.1609/aaai.v32i1.11934.

[4] F. Alam and G. Riccardi, "Predicting personality traits using multimodal information," in Proceedings of the 2014 ACM multimedia workshop on computational personality recognition, 2014, pp. 15–18. doi: 10.1145/2663204.2663251.

[5] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in Natural language processing and Chinese computing: 4th CCF conference, NLPCC 2015, Nanchang, China, October 9-13, 2015, proceedings 4. Springer, 2015, pp. 159–167. doi: 10.1007/978-3-319-25207-0_14.

[6] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 2: Short papers), 2018, pp. 606–611. doi: 10.18653/v1/P18-2096.

[7] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," arXiv, 2017. unpublished. Available: http://arxiv.org/abs/1707.07250.

[8] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," arXiv, 2021. unpublished. Available: http://arxiv.org/abs/2109.00412.

[9] H. Akbari, L. Yuan, R. Qian, W. H. Chuang, S. F. Chang, Y. Cui, et al., "Vatt: transformers for multimodal self-supervised learning from raw video, audio and text," Advances in Neural Information Processing Systems, vol. 34, pp. 24206–24221, 2021.

[10] B. Dufumier, P. Gori, J. Victor, A. Grigis, M. Wessa, P. Brambilla, et al., "Contrastive learning with continuous proxy meta-data for 3d mri classification," in Medical image computing and computer assisted intervention – MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part II. Springer, 2021, pp. 58–68. doi: 10.1007/978-3-030-87196-3_6.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, vol. 139, 2021, pp. 8748–8763.

[12] G. Hu, T. E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "Unimse: towards unified multimodal sentiment analysis and emotion recognition," arXiv, 2022. unpublished. Available: http://arxiv.org/abs/2211.11256.

[13] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi, "Rank-n-contrast: learning continuous representations for regression," Advances in Neural Information Processing Systems, vol. 36, 2024.

[14] R. Li, G. Gao, S. Song, and K. Xiao, "Multimodal sentiment analysis method based on improved FGM-CM-BERT model," Journal of Hebei University (Natural Science Edition), vol. 45, no. 2, pp. 192–203, 2025. [Journal of Hebei University(Natural Science Edition), vol. 45, no. 2, pp. 192–203, 2025.]

[15] P. Wang, Q. Zhou, Y. Wu, et al., "DLF: disentangled-language-focused multimodal sentiment analysis," in Proceedings of the AAAI conference on artificial intelligence, vol. 39, no. 20, 2025, pp. 21180–21188.

[16] R. Zeng, C. Wang, and Q. Chen, "Comparative study on stages and models of online public opinion dissemination," Journal of Information, vol. 33, no. 5, pp. 119–124, 2014. [Journal of Information, vol. 33, no. 5, pp. 119–124, 2014.]

[17] J. Chen, Y. Wang, and H. Nie, "Research on evolution analysis model of short video public opinion in emergency events," Journal of Information Resource Management, vol. 12, no. 3, pp. 152–164, 2022. doi: 10.13365/j.jirm.2022.03.152. [Journal of Information Resource Management, vol. 12, no. 3, pp. 152–164, 2022. doi: 10.13365/j.jirm.2022.03.152.]

[18] H. Wei, H. Zhu, J. Wei, et al., "Research on internet public opinion evolution based on short video networks," Data Analysis and Knowledge Discovery, vol. 8, no. 5, pp. 113–126, 2024. [Data Analysis and Knowledge Discovery, vol. 8, no. 5, pp. 113–126, 2024.]

[19] A. Zadeh, R. Zellers, E. Pincus, et al., "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," IEEE Intelligent Systems, vol. 31, no. 6, pp. 82–88, 2016. doi: 10.1109/MIS.2016.94.

[20] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML), 2018, pp. 11–19. doi: 10.18653/v1/W18-3302.

[21] J. Williams, R. Comanescu, O. Radu, and L. Tian, "DNN multimodal fusion techniques for predicting video sentiment," in Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML), 2018, pp. 64–72. doi: 10.18653/v1/W18-3308.

[22] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," arXiv, 2018. unpublished. Available: http://arxiv.org/abs/1806.00064.

[23] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in Proceedings of the AAAI conference on artificial intelligence. 2018;32(1). doi:10.1609/aaai.v32i1.11934.

[24] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: Long papers), 2018, pp. 2236–2246. doi: 10.18653/v1/P18-1208.

[25] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proceedings of the 57th annual meeting of the Association for Computational Linguistics, 2019, pp. 6558–6569. doi: 10.18653/v1/P19-1656.

[26] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2554–2562. doi: 10.1109/CVPR46437.2021.00258.

[27] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: modality-invariant and-specific representations for multimodal sentiment analysis," in Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1122–1131. doi: 10.1145/3394171.3413678.

[28] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L. P. Morency, et al., "Integrating multimodal information in large pretrained transformers," in Proceedings of the 58th annual meeting of the Association for Computational Linguistics, 2020, pp. 2359–2369. doi: 10.18653/v1/2020.acl-main.214.

[29] Y. Li, Y. Wang, and Z. Cui, "Decoupled multi-modal distilling for emotion recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 6631–6640. doi: 10.1109/CVPR52729.2023.00641.

[30] Z. Guo, T. Jin, J. Chen, J. Zhang, Y. Zhang, and H. Wang, "Classifier-guided gradient modulation for enhanced multimodal learning," Advances in Neural Information Processing Systems, vol. 37, pp. 133328–133344, 2024.

[31] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," IEEE Transactions on Affective Computing, vol. 15, no. 1, pp. 309–325, 2024.

[32] J. Lin, Y. Wang, Y. Xu, Y. Wang, and Y. Yang, "Semi-IIN: Semi-supervised intra-inter modal interaction learning network for multimodal sentiment analysis," in Proceedings of the AAAI Conference on Artificial Intelligence, 2025, vol. 39, no. 2, pp. 1411–1419.

[33] J. Peng, Y. He, Y. Chang, Z. Li, and L. Zhao, "A social media dataset and H-GNN-based contrastive learning scheme for multimodal sentiment analysis," Applied Sciences, vol. 15, p. 636, 2025.

[34] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of spanish online videos," IEEE Intelligent Systems, 28(3):38–45, 2013.

[35] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency, "Deep multimodal fusion for persuasiveness prediction," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 284–288, 2016.

[36] L. Chen and G. Zhu, "Self-supervised contrastive learning for itinerary recommendation," Expert Systems with Applications, 268:126246, 2025.

[37] Y. Shou, X. Cao, and D. Meng, "Spegcl: Self-supervised graph spectrum contrastive learning without positive samples," IEEE Transactions on Neural Networks and Learning Systems, in press.

[38] L. Zhao, Y. Zhang, J. Duan, et al, "Cross-supervised contrastive learning domain adaptation network for steel defect segmentation," Advanced Engineering Informatics, 64:102964, 2025.

[39] S. Yang, Y. Yang, D. Zhao, et al, "Dynamic malware detection based on supervised contrastive learning," Computers and Electrical Engineering, 123:110108, 2025.

[40] I. Nesteruk, "General SIR model for visible and hidden epidemic dynamics," Frontiers in Artificial Intelligence, 8:1559880, 2025.

[41] M. Corteel, "Shaping epidemic dynamics: An historical epistemology study of the SIR model," History of the Human Sciences, in press.