News Event Source Analysis with Multimodal Transformer Integration

Yao Li[1], Shaohua Wei[1], Tian Bao[1], Wei Gao[2]
[1]Department of Journalism and Communication, Shanghai University of Sport, Shanghai 200438, China
[2]The Shanghai Pudong Micro-Hotspot Big Data Research Institute, Shanghai 200100, China
Corresponding author: Shaohua Wei (e-mail: weishaohua2024@163.com)

ABSTRACT To address the limitations of current news event tracing methods, including reliance on single-modal information, fragmented evidence, and insufficient temporal dynamic modeling, this paper proposes a hybrid framework that integrates a multimodal Transformer with a Temporal Graph Attention Network (T-GAT) to improve the accuracy and robustness of event source tracing. The framework first extracts enhanced representations from text, images, and audio, and employs a two-stage cross-attention mechanism to achieve deep cross-modal fusion and generate unified event embeddings. Subsequently, temporal embeddings and multi-source propagation relationships are incorporated into T-GAT for structured reasoning to identify event origins and reconstruct dissemination paths. From an engineering perspective, the proposed multimodal fusion and temporal reasoning strategy also provides a generalizable methodology for heterogeneous information analysis in intelligent communication environments, offering potential reference value for electromagnetic information processing and propagation-aware decision support. Furthermore, the framework demonstrates transferable applicability to industrial scenarios. Experimental results show that the proposed method achieves a Top-1 tracing accuracy of 0.92 and a Path-IoU of 0.66 while maintaining superior robustness in propagation path reconstruction. Although the model requires higher computational resources than lightweight baselines, its advantages in tracing accuracy and interpretability provide a reliable solution for complex event attribution tasks.

INDEX TERMS multimodal transformer, temporal graph attention network, news event tracing, electromagnetic information processing, heterogeneous data fusion

I. INTRODUCTION

WITH the rapid development of social platforms and online media, information spreads rapidly across platforms and languages in a short period of time. A comparable challenge exists in the modern, globalized industry, where issues like a critical defect in a raw material batch or a machine malfunction can rapidly propagate their consequences across interconnected digital systems, quickly impacting downstream processes, different factories, and global supply chain partners who share real-time production data [1-3].

False information, quoting out of context, and multiple reposts make it increasingly difficult to accurately identify the source of information and restore the dissemination path, which has a direct impact on fact-checking, public opinion management, and legal accountability [4-6]. Existing deep learning attribution methods often rely on a single modality (primarily text) or only utilize the dissemination structure, generally ignoring multimodal evidence such as images and audio and the temporal dynamics of dissemination. As a result, they show significant deficiencies in robustness and interpretability in the face of metadata tampering, cross-platform reposting, and multimodal forgery. There is an urgent need for an attribution method that simultaneously integrates multimodal evidence and explicitly models temporal dynamics to improve localization accuracy and provide a verifiable chain of evidence.

In recent years, multimodal representation and deep learning have made significant progress in fake news detection and news dissemination analysis. Uppada S K et al. achieved an accuracy of 91.94% on the Fakeddit dataset with an image and text-based framework, proving the value of visual features [7]; Palani B fused the text features of BERT with the visual features extracted by CapsNet (Capsule Networks), significantly improving the classification performance [8]; Kishore V achieved an accuracy of 96.51% on multimodal fake news classification through optimal feature fusion and improved BiLSTM, demonstrating the potential of feature optimization and sequence modeling [9]; Guo Y's MANN (mutual attention neural network) uses the mutual attention mechanism to strengthen inter-modal correction, outperforming various SOTA (State of the Art) methods on Weibo data [10]; Kumari R integrates novelty detection and sentiment prediction of supervised contrastive learning into a multimodal framework, better identifying multimedia information that is misused in new contexts [11]; Zhang G's

SceneFND (Scene Fake News Detection) model explicitly models scene context and text representation, and The baseline is improved by approximately 3.48% and 3.32% on PolitiFact and GossipCop, respectively, proving that scene context is helpful for discrimination [12].

Existing cross-modal news event attribution methods generally suffer from insufficient accuracy, stemming from the limited scope of single-modal evidence and the lack of dynamic propagation modeling. Similarly, in industrial scenarios such as product defect attribution faces significant challenges due to fragmented data sources (e.g., isolated fabric images, sensor logs, and batch records) and a lack of time-series dependency tracking. To address these limitations, this paper proposes an innovative hybrid framework that integrates a multimodal Transformer and a temporal graph attention network (T-GAT). This framework first extracts enhanced features through modality-specific encoders (BERT combined with hierarchical attention for text, ViT integrated forensic features for images, and Wav2Vec2 with cross-modal anchor boxes for audio). Next, a two-stage cross-attention mechanism is designed to achieve cross-modal alignment and generate a unified joint representation. Finally, the T-GAT module integrates modal features, temporal embeddings, and multi-source edge information to perform interpretable time-aware graph reasoning, seamlessly coupling micro-content analysis with macro-propagation dynamics to achieve high-precision source localization and robust path reconstruction. This method not only significantly improves the performance of news event attribution (Top-1 attribution accuracy reaches 0.92, an improvement of 0.09 compared to the VisualBERT baseline), but also provides a transferable solution for industrial applications such as defect root cause analysis through modular feature mapping and the robust inference mechanism of T-GAT, effectively bridging the multi-source data gap and strengthening temporal causal inference.

II. TRACING FRAMEWORK INTEGRATING MULTIMODAL TRANSFORMER AND T-GAT

A. TEXT SEMANTICS AND STANCE EXTRACTION MODULE

B. TRANSFORMER-BASED CONTEXT ENCODING

The text is first normalized (denoising, entity decoding, and character set unification), and then segmented using the BERT subtokenizer WordPiece to obtain a token sequence [13]. To cope with extremely long news and cross-segment dependencies in context, an overlapping segmentation strategy is adopted to segment the sequence, with each segment length not exceeding 512.

For each sub-sequence, a pre-trained BERT encoder in a deep learning method is used to obtain the token representation matrix within the segment. In this paper, a pre-trained BERT model is used as the text encoder instance. Its specific structure follows the original BERT design, including a word segmentation mechanism based on WordPiece, an input representation consisting of word embeddings, segment embeddings, and positional embeddings, and a deep context encoding network formed by stacking multiple layers (L=12/24) of Transformer Blocks. It is pre-trained using the MLM/NSP objective. The above embedding construction and multi-layer stacking process are fully preserved in the engineering implementation. However, in order to avoid introducing implementation details that are not directly related to the contribution of this paper into the method description, this paper only abstracts the general operation form of a single-layer Transformer at the formula level to emphasize its functional role in cross-modal fusion and subsequent graph inference. In this paper, BERT is regarded as a specific, pre-trained and initialized Transformer encoder instance, and Equation (1) is used to uniformly describe the computation mechanism of its internal layers, rather than replacing or generalizing the complete model definition of BERT. The single-layer operation definition of BERT is shown in formula (1) [14,15].

$$\begin{aligned} Q = T_e W_Q, \quad K = T_e W_K, \quad V = T_e W_V, \\ \mathrm{head}_i(T_e) = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i, \\ \mathrm{MH}(T_e) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_\beta) W_O, \\ \widetilde{T}_e = \mathrm{LN}(T_e + \mathrm{MH}(T_e)), \\ T'_e = \mathrm{LN}\left(\widetilde{T}_e + \mathrm{FFN}(\widetilde{T}_e)\right). \end{aligned} \tag{1}$$

In formula (1), W_Q, W_K, and W_V represent the weight matrices corresponding to the query, key, and value, W_O represents the projection matrix, and β represents the number of attention heads. FFN (Feed-Forward Neural Network) represents a feed-forward neural network, LN represents LayerNorm, and T_e represents the intra-segment token representation matrix.

C. HIERARCHICAL AGGREGATION AND ATTENTION-WEIGHTED POOLING

To obtain document-level representation while taking into account inter-segment context, a two-layer hierarchical aggregation is adopted. First, internal pooling is performed on each segment to obtain a segment vector. Then, self-attention aggregation is performed on the segment sequence to obtain a document vector sequence. Finally, attention weighting is used to obtain a document fusion representation.

Intra-segment aggregation uses attention pooling with learning. The attention weight calculation and segment vector definition are shown in formula (2) [16-18].

$$\begin{aligned} \alpha_{(j,t)} &= \frac{\exp\left(q_p^\top \tanh(W_a h_{(j,t)} + b_a)\right)}{\sum_{t'} \exp\left(q_p^\top \tanh(W_a h_{(j,t')} + b_a)\right)}, \\ s_j &= \sum_{t=1}^{L_j} \alpha_{(j,t)} h_{(j,t)}. \end{aligned} \tag{2}$$

In formula (2), $h_{(j,t)}$ represents Word-level hidden states output by the text encoder, and $q_p^\top$ represents the pooled query vector. W_a and b_a represent the weight matrix and bias in the attention weight calculation, respectively. s_j represents the segment vector.

Segmented attention aggregation is applied to segmented vector sequences to obtain segmented context-enhanced representations. The final document-level joint representation is obtained through weighted aggregation, as shown in formula (3).

$$\begin{aligned} h_{\text{text}} &= \sum_{j=1}^{J} \gamma_j \bar{s}_j, \\ \gamma_j &= \frac{\exp\left(\delta^\top \tanh(W_b \bar{s}_j + b_b)\right)}{\sum_{j'} \exp(\delta^\top \tanh(W_b \bar{s}_{j'} + b_b))}. \end{aligned} \tag{3}$$

In formula (3), $\bar{s}_j$ represents the segment-level context-enhanced representation, and $\delta^\top$ represents the learnable parameters.

To support subsequent temporal graph reasoning and cross-modal alignment, the text representation explicitly incorporates release time and platform metadata. It defines a temporal encoding function and platform/media embeddings, projects them, concatenates them with the text representation, and then projects them back to a unified dimension, as shown in formula (4).

$$\bar{h}_{\text{text}} = W_p[h_{\text{text}} \| \tau(t_{\text{pub}}) \| m_o] + b_p \tag{4}$$

In formula (4), $\tau(t_{\text{pub}})$ represents the temporal encoding function, and m_o represents the platform embedding.

D. IMAGE EVIDENCE ENCODING MODULE

E. FORENSIC FEATURE EXTRACTION

The input color raw image is first normalized and geometrically preprocessed. It is then cropped or scaled to a fixed resolution and pixel-normalized. Several image forensic quantitative features are computed in parallel, including ELA (Error Level Analysis) residuals, noise residuals, and PRNU (Photo Response Non-Uniformity) fingerprint estimation. DCT (Discrete Cosine Transform)/JPEG (Joint Photographic Experts Group) quantitative inconsistency statistics serve as external evidence for subsequent graph reasoning.

Finally, the forensic feature sets are concatenated. In the ELA residual, the normalized image is recompressed to JPEG quality, the residual is calculated, and global second-order statistics (mean, variance, and high-order moments) are used to obtain the eigenvector components.

F. VIT ENCODER STACKING AND FINE-GRAINED REPRESENTATION OUTPUT

For the normalized image, it is divided into non-overlapping patches, and all patches are linearly mapped and positionally encoded, as shown in (5).

$$\begin{aligned} z_k^{(0)} &= W_e d_k + b_e, \\ Z^{(0)} &= [z_0^{(0)}; z_1^{(0)}; \ldots; z_P^{(0)}] + E_{\text{pos}}. \end{aligned} \tag{5}$$

In formula (5), E_{pos} represents the learnable position encoding matrix, W_e and b_e represent the corresponding projection matrix and bias, and d_k represents the pixel representation of the patch flattened into a vector.

Let the Transformer layer index be L. Each layer performs multi-head self-attention and feed-forward transformation on the input, as shown in formula (6) [19].

$$\begin{aligned} &Q^{(l)} = Z^{(l-1)} W_Q^{(l)}, \quad K^{(l)} = Z^{(l-1)} W_K^{(l)}, \quad V^{(l)} = Z^{(l-1)} W_V^{(l)}, \\ &A^{(l)} = \text{softmax}\left(\frac{Q^{(l)} K^{(l)\top}}{\sqrt{d_k}}\right), \\ &\text{SA}^{(l)} = A^{(l)} V^{(l)}, \\ &\widetilde{Z}^{(l)} = \text{LN}\left(Z^{(l-1)} + \text{SA}^{(l)} W_O^{(l)}\right), \\ &Z^{(l)} = \text{LN}\left(\widetilde{Z}^{(l)} + \text{FFN}^{(l)}(\widetilde{Z}^{(l)})\right). \end{aligned} \tag{6}$$

In formula (6), $Z^{(l-1)}$ represents the input data, $Z^{(l)}$ represents the final output, and $A^{(l)}$ represents the attention matrix.

To ensure compatibility between visual features and subsequent multimodal fusion, a learnable projection head is introduced based on the class labels output by ViT. This design maps the high-dimensional features of the original ViT output to a unified semantic space through linear transformation and layer normalization operations, effectively mitigating the distributional differences between visual features and text/audio features. This adaptation mechanism preserves the high-level semantic information extracted by ViT while enhancing the adaptability of features to multimodal alignment tasks, providing a more consistent representational foundation for subsequent cross-modal interactions.Taking the linear transformation of the class token and performing normalization, as shown in formula (7).

$$h_{\text{img}} = \text{LN}(W_h z_0^{(L)} + b_h) \tag{7}$$

In formula (7), h_{img} is the global visual representation.

Now, the previously extracted forensic vector is projected into the same dimension and concatenated with the global visual representation to form an extended visual vector, as shown in formula (8).

$$\bar{h}_{\text{img}} = W_\zeta[h_{\text{img}} \| \zeta_{\text{for}}] + b_\zeta \tag{8}$$

In formula (8), ζ_{for} represents the extracted forensic vector, and $\bar{h}_{\text{img}}$ represents the final extended visual vector. The forensic vector here is a composite representation constructed by fusing multiple image forensic features, including global second-order statistics (mean, variance, and higher-order moments) of ELA residuals, noise distribution features, PRNU fingerprint features, and DCT/JPEG compression consistency features. These features together

constitute a quantitative characterization of the reliability of the image source and potential tampering traces, rather than a single type of feature extraction result.

G. AUDIO EVIDENCE CODING MODULE

The input audio files are first uniformly resampled to 16kHz. Long files are then segmented into overlapping segments using fixed-length windows. Endpoint detection is performed on each segment, and silence segments are removed, while speech/ambient sound segments are retained to reduce noise interference.

The original waveform of each segment is input into the front-end convolutional feature encoder of the Wav2Vec2 model, which consists of a multi-layer temporal convolutional network responsible for extracting basic acoustic features [20,21]. This paper uses the frame-level latent representation output by this convolutional encoder as the basis for audio features, instead of using its subsequent Transformer module, to maintain consistency with the computational granularity of the text and image modal feature extraction stages. The frame-level latent representation is shown in Equation (9).

$$e_\iota = \theta\big(\hat{x}[n_k^{\text{start}} : n_k^{\text{end}}]\big)_\iota \tag{9}$$

In formula (9), ι represents the frame index, k represents the segment index, and $\hat{x}$ represents the audio signal.

To support temporal alignment with text, ASR (Automatic Speech Recognition) is used to map segment-level timestamps to text token indices to generate a set of cross-modal anchor points. In the subsequent multimodal Transformer, these anchor points are used to construct the audio-to-text attention mask and the initial similarity edge weights to ensure that temporal consistency evidence can be aggregated first. The final output is L2 normalized according to the interface convention, as shown in formula (10).

$$\bar{h}_{\text{audio}} = \frac{h_{\text{audio}}}{\|h_{\text{audio}}\|} \tag{10}$$

In formula (10), $\bar{h}_{\text{audio}}$ represents the normalized audio features, and h_{audio} represents the global audio vector.

The multimodal Transformer and T-GAT hybrid framework is shown in Figure 1.

Figure 1 shows the hybrid architecture of this paper, which integrates a multimodal Transformer and T-GAT.

The system receives multi-source information, including text, images, audio, and metadata. It encodes unimodal features using BERT, ViT, and Wav2Vec2, respectively. After dimensionality alignment at the modal projection layer, the representation is fed into a multimodal Transformer for cross-modal interaction and fusion, resulting in a unified joint representation. This representation is used for text classification and sentiment recognition tasks, extracting the factual stance and emotional characteristics of news. It is also injected into the propagation graph as node features, jointly constructing an event propagation graph with metadata and modal evidence. T-GAT then performs graph attention reasoning based on temporal information and propagation relationships, outputting event sources, edge credibility, and propagation paths.

In this framework, "enhanced features" refers to semantically strengthened representations of raw modal data through domain-specific deep processing and feature engineering. For text modality, a BERT pre-trained model combined with a hierarchical attention aggregation mechanism is used to decompose long news articles into context-coherent segments, while integrating publication timestamps and platform metadata as structured supplements. For image modality, the enhancement process integrates high-level semantic features extracted by Vision Transformer with low-level clues generated by professional image forensics techniques, including anti-counterfeiting indicators such as error level analysis residuals, noise distribution features, and non-uniform light response fingerprints. For audio modality, enhanced features originate from acoustic representations extracted by Wav2Vec2, combined with endpoint detection noise reduction and cross-modal time anchors established through automatic speech recognition to ensure accurate temporal alignment of audio and video content. These targeted enhancement strategies go beyond basic feature extraction, significantly improving the discriminative power and source traceability reliability of the raw data by injecting domain knowledge, temporal context, and cross-modal associations, laying a solid foundation for subsequent multimodal fusion.

H. JOINT REPRESENTATION LEARNING WITH MULTIMODAL TRANSFORMERS

I. CONSTRUCTING CROSS-MODAL TOKEN SEQUENCE AND POSITION MAPPING

The output of each mode is first linearly mapped and normalized, as shown in (11).

$$\begin{aligned}
\tilde{h}_{\text{text}} &= \frac{W_{\text{text}}\bar{h}_{\text{text}} + b_{\text{text}}}{\|W_{\text{text}}\bar{h}_{\text{text}} + b_{\text{text}}\|}, \\
\tilde{h}_{\text{img}} &= \frac{W_{\text{img}}\bar{h}_{\text{img}} + b_{\text{img}}}{\|W_{\text{img}}\bar{h}_{\text{img}} + b_{\text{img}}\|}, \\
\tilde{h}_{\text{audio}} &= \frac{W_{\text{audio}}\bar{h}_{\text{audio}} + b_{\text{audio}}}{\|W_{\text{audio}}\bar{h}_{\text{audio}} + b_{\text{audio}}\|}.
\end{aligned} \tag{11}$$

In formula (11), W_{text}, W_{img}, and W_{audio} represent the corresponding linear projection matrices, and b_{text}, b_{img}, and b_{audio} represent the corresponding biases.

The release time embedding and platform embedding are projected and incorporated into each modality vector, as shown in formula (12). LayerNorm is then used to normalize the time and source biases, encoding them into the semantic vector of each modality.

$$\hat{h}_{m_o} = W_{m_o}^{(\tau)}\tau(t_{\text{pub}}) + W_{m_o}^{(p)} + \tilde{h}_{m_o} \tag{12}$$

In formula (12), m_o represents the platform embedding of text, image, and audio modalities.

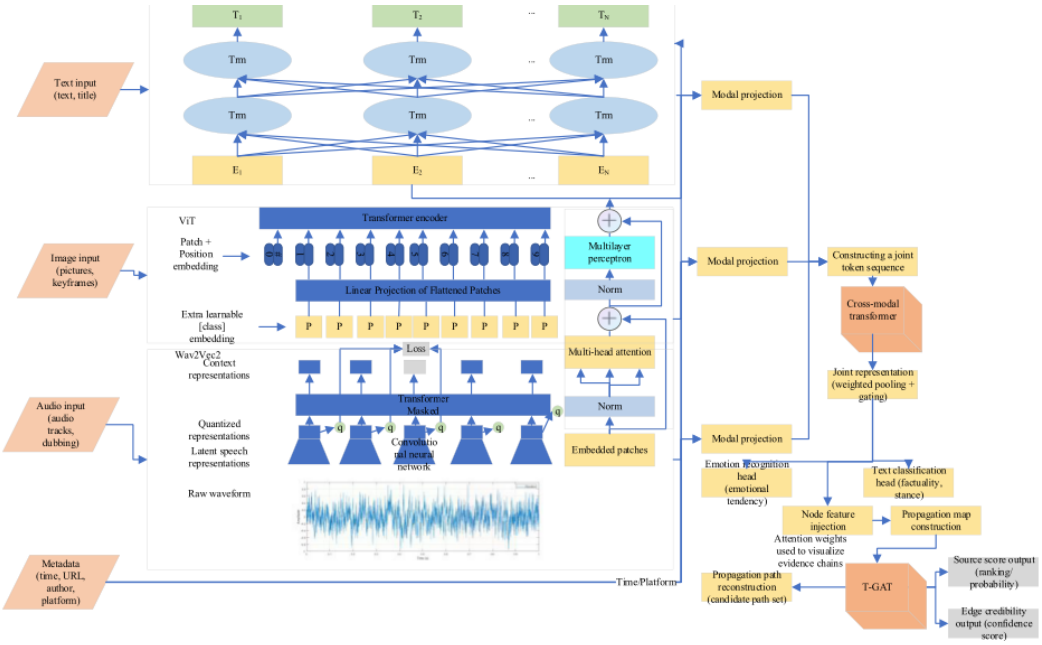

FIGURE 1. Multimodal Transformer and T-GAT hybrid framework

For fine-grained alignment requirements, retain the text token list, image patch token list, and audio segment token list, and use linear layers to map them to the common dimension, respectively. Adding the modality type embedding and position/time embedding to the token to obtain the joint input sequence, as shown in formula (13).

$$X = [x_1, \ldots, x_N], \quad x_n = v_n + E_{\text{type}}(m_{o_n}) + E_{\text{pos}}(\text{pos}_n) \tag{13}$$

In formula (13), v_n represents the mapped token vector, E_{pos} represents the position/time embedding, and E_{type} represents the modality type embedding.

J. CROSS-MODAL TRANSFORMER FUSION AND OUTPUT LAYER DESIGN

This paper implements multi-head attention on input features using a fusion encoder based on multi-head self-attention, achieving modality fusion [22-24]. The attention weight calculation is shown in formula (14).

$$A(t) = \text{softmax}\left(\frac{Q(t)K(t)^\top}{\sqrt{d_k}} + F(t)\right) \tag{14}$$

In formula (14), $F(t)$ represents the attention bias matrix under modality and time conditions, which is used to improve the discriminative power of cross-modal alignment. To emphasize text-centric traceability clues, a two-stage attention is designed: the first stage is full-sequence self-attention; the second stage performs text-centric cross-attention on text tokens, where the query comes from the text token and the key value comes from the contextual representation of image/audio tokens, as shown in formula (15). The text representation is injected in a residual manner, allowing the text to "actively query" visual and auditory evidence during fusion.

In this framework, the two-stage cross-attention mechanism is executed sequentially. The first stage is a full-modal self-attention stage, which concatenates all tokens from text, images, and audio into a unified sequence. Global interaction is achieved through a standard multi-head self-attention mechanism, enabling initial fusion of features across modalities and capturing fundamental cross-modal associations. The second stage is a text-guided cross-attention stage, specifically strengthening the dominant role of text over other modalities: the text tokens output from the first stage are used as queries, while the image and audio tokens output from the first stage are used as keys and values. This design allows text content to actively "query" visual and auditory evidence, selectively extracting multimodal information most relevant to the text's semantics. Residual connections are used between the two stages to preserve the core information of the original text features, ensuring that source tracing clues are not diluted. This sequential architecture guarantees full-modal interaction in the initial stage while highlighting the central role of text in the news source tracing task in the second stage, enabling the model to prioritize the selection and verification of multimodal evidence based on semantic content.

$$\begin{aligned} Q_t &= X_{\text{text}} W_Q^t, \\ K_{ia} &= \text{Concat}(X_{\text{img}}, X_{\text{audio}}) W_K^{ia}, \\ V_{ia} &= \text{Concat}(X_{\text{img}}, X_{\text{audio}}) W_V^{ia}, \\ C_t &= \text{softmax}\left(\frac{Q_t K_{ia}^\top}{\sqrt{d_k}} + F_t\right) V_{ia}. \end{aligned} \tag{15}$$

After completing multi-layer fusion, the final sequence XLf is obtained. To obtain event-level joint representa-

tion, modality-aware gated pooling is applied. The pooling weight is calculated for each token to obtain the joint representation, as shown in (16).

$$\hat{\alpha}_n = \frac{\exp\left(w^\top \tanh(W_p x_n + b_p) + \lambda_{m_{o_n}}\right)}{\sum_{n'} \exp\left(w^\top \tanh(W_p x_{n'} + b_p) + \lambda_{m_{o_{n'}}}\right)},$$
$$h_{\text{multi}} = \sum_{n=1}^{N} \hat{\alpha}_n x_n. \tag{16}$$

In formula (16), $\lambda_{m_{o_n}}$ represents the modal bias term, w, W_p, and b_p represent learnable parameters, and h_{multi} represents the joint representation of the weighted sum of the modalities.

Simultaneously, the modal gating vectors (corresponding to text, img, and audio, respectively) are calculated to evaluate the relative contribution of each modality to the final decision, and h_{multi} is calibrated according to the gating, as shown in formula (17).

$$g = \sigma(W_g[\bar{h}_{\text{text}} \| \bar{h}_{\text{img}} \| \bar{h}_{\text{audio}}] + b_g),$$
$$\tilde{h}_{\text{multi}} = \text{LN}(h_{\text{multi}} \odot g_{\text{exp}}). \tag{17}$$

In formula (17), g represents the three-dimensional modal gating vector $[g_{\text{text}}, g_{\text{img}}, g_{\text{audio}}]$, where each component corresponds to the weight coefficient of the text, image, and audio modalities, respectively.

This vector is calculated by fusing the global representations of each modality and is a global gating parameter related to the event-level representation, shared by all tags of the current event. g_{exp} represents the vector obtained by repeatedly expanding the 3D gated vector according to the number of markers for each modality: for the text modality with L_{text} markers, g_{text} is repeated L_{text} times; for the image modality with L_{img} markers, g_{img} is repeated L_{img} times; and for the audio modality with L_{audio} markers, g_{audio} is repeated L_{audio} times. The expanded g_{exp} has the same length as the marker sequence and is used for element-wise multiplication with h_{multi}. This enables the broadcasting of modal gating vectors to the token dimension and their element-wise application to calibrate the fused representation. Layer normalization is then used to stabilize the scale variance introduced by the gating operation, where h_{multi} represents the final calibrated fused vector.

In this paper, the gating design adopts the order of applying modal weights first and then performing layer normalization, mainly because: (1) the gating operation will change the scale distribution of the original fused representation, which may lead to a significant amplification or reduction of the values of some dimensions; (2) applying layer normalization (LN) after gating can effectively stabilize this scale change and ensure that the representation processed later has stable statistical properties; (3) if LN is performed first and then gating is applied, the established normalization properties will be destroyed, resulting in unstable training.

Based on the fusion vector, two classification heads (text classification and sentiment recognition) are constructed, as shown in formula (18).

$$\hat{y}_{\text{class}} = \text{Softmax}(W_c \tilde{h}_{\text{multi}} + b_{\text{class}}),$$
$$\hat{y}_{\text{sent}} = \text{Softmax}(W_{\text{sent}} \tilde{h}_{\text{multi}} + b_{\text{sent}}). \tag{18}$$

In formula (18), $\hat{y}_{\text{class}}$ and $\hat{y}_{\text{sent}}$ represent the text classification probability distribution and sentiment probability distribution, respectively. These tasks serve as auxiliary supervision signals and have a dual role:

(1) Enhancing the representation ability of the multimodal Transformer through multi-task learning. Text classification (factual/stance-based) and sentiment recognition tasks provide additional supervision signals for the model, prompting the multimodal Transformer to learn more discriminative joint representations, thereby improving the performance of subsequent source tracing tasks; (2) Extracting key semantic features. The factual stance and sentiment of news are important factors affecting the propagation path (e.g., negative sentiment content tends to spread faster). These features are implicitly encoded in the joint representation, providing supplementary information for the T-GAT propagation path reconstruction.

K. TEMPORAL PROPAGATION GRAPH REASONING BASED ON T-GAT

L. PROPAGATION GRAPH CONSTRUCTION

For each node, each news item, post, or URL (Uniform Resource Identifier) is considered a basic node, and account/media nodes are added as extensions. Each node is accompanied by a multimodal vector and metadata vectors (platform embedding, author reputation, language tags, etc.). For each edge, three types of sources are constructed and merged. Explicit edges include parsed citations, forwarding, and reposting.

For similarity edges, if the modal similarity function exceeds a threshold, an undirected or directed similarity edge is added. For temporal proximity edges, edges are constructed from the N closest nodes to the same event based on publication time to ensure connectivity. In constructing the diffusion graph, edge feature vectors include multimodal weighted cosine similarity, local visual/audio similarity, temporal difference, and one-hot citation type.

The unified joint representation generated by the multimodal Transformer is embedded into the T-GAT propagation graph through a dual integration mechanism. This representation serves as the core initial node feature vector, directly replacing the simple embedding in traditional graph networks and providing a rich semantic foundation for each news node. Secondly, these joint representations are feature-concatenated with metadata (including publication timestamps, platform type, and author reputation) and fused through a lightweight projection layer to form enhanced node representations.

During T-GAT's attention computation, these enhanced representations serve both as feature inputs to the nodes themselves and participate in the dynamic calculation of edge weights: when evaluating the propagation credibility between two nodes, the model calculates the cosine similarity between their joint representations and combines this value with a time decay factor and platform compatibility metrics to form initial edge weights. More importantly, within the attention mechanism, the joint representations of nodes generate query and key vectors through a learnable mapping function, enabling T-GAT to adaptively weight neighbor information based on semantic relevance rather than just topological structure. This deep integration ensures that multimodal semantic information not only initializes the graph structure but also continuously guides the entire time-aware reasoning process, allowing source tracing decisions to benefit from both content semantics and propagation dynamics.

The constructed diffusion graph is shown in Figure 2.

Figure 2 illustrates the multi-platform dissemination paths and temporal evolution of a news event. It can be seen that the event is initially published by "News_A" on a news website and subsequently spreads rapidly through social media platforms (such as "Twitter_B" and "Blog_D") and forum platforms (such as "Forum_C"), forming multiple parallel dissemination chains. The edge weights between nodes reflect the combined influence of content similarity and time decay. Thicker edges indicate information dissemination with high similarity over a short period of time, while dashed edges indicate implicit citations or forwarding relationships. Nodes across platforms exhibit a distinct hierarchical layout, with cross-propagation between different platforms, indicating complex interactions between information media. Furthermore, some nodes (such as "News_E" to "Twitter_K" and then to "News_L") exhibit secondary or even tertiary diffusion, demonstrating the event's sustained spread and multimodal influence. Overall, the node and edge structure in the figure reveals the temporal dynamics of news event dissemination, cross-platform diffusion paths, and information strength relationships, providing a visual reference for source tracing analysis.

M. SEQUENCE GRAPH ATTENTION REASONING MODULE

This paper is based on the deep learning GAT (Graph Attention Networks) model. In order to apply time priority in attention, a time embedding function is defined and a learnable relative time embedding is adopted, as shown in (19) [25].

$$\tau(\Delta t) = W_\tau \phi(\Delta t),$$
$$\phi(\Delta t) = \left[\sin\left(\frac{\Delta t}{10000^{2k/d_t}}\right), \cos\left(\frac{\Delta t}{10000^{2k/d_t}}\right)\right]_{k=0}^{d_t/2-1}. \tag{19}$$

In formula (19), $\tau(\Delta t)$ represents the temporal embedding function.

First, a linear transformation is performed on the node representation. For each edge, the unnormalized attention score is calculated as shown in (20). Then, normalization is performed.

$$\bar{\alpha}_{ij}^{(l)} = \mathrm{LeakyReLU}\left(\mathbf{o}^{(l)\top}[\tilde{h}_i^{(l)} \| \tilde{h}_j^{(l)} \| \phi_e(\hat{e}_{ij}) \| \tau(\Delta t_{ij})]\right) \tag{20}$$

In formula (20), $\tilde{h}_i^{(l)}$ represents the node representation after linear transformation, $\hat{e}_{ij}$ represents the unnormalized attention score, and $\mathbf{o}^{(l)\top}$ represents the learnable vector.

In a multi-head setting, parallel attention heads are used and the results are averaged, as shown in formula (21).

$$\mathrm{h}_i^{(l+1)} = \sigma\left(\Big\|_{k=1}^{H} \sum_{j \square N(i)} \alpha_{ij}^{(l,k)} U^{(l,k)} \mathrm{h}_j^{(l)}\right) + Res_i^{(l)} \tag{21}$$

In formula (21), $U^{(l,k)}$ represents the message transformation matrix for each head, σ represents the nonlinear ReLU (Rectified Linear Unit), and $\mathrm{Res}_i^{(l)}$ represents the residual term followed by LayerNorm. In the final layer, the aggregated node representation and the original edge features are used to predict edge credibility, as shown in formula (22).

$$\rho_{ij} = \sigma\left(w_c^\top [h_i^{(L)} \| h_j^{(L)} \| \phi_e(e_{ij})] + b_c\right) \tag{22}$$

In formula (22), ρ_{ij} represents edge credibility.

For the same event subgraph, the source score is calculated based on the final node representation and structured context, as shown in formula (23).

$$\begin{cases} \pi_i = \dfrac{\exp(Q_i)}{\sum_{j \in V_{N_e}} \exp(Q_j)}, \\ Q_i = \varpi_{1Q}^\top h_i^{(L)} + \varpi_{2Q}^\top \left(\sum_{j \in N(i)} \rho_{ij} \cdot h_j^{(L)}\right). \end{cases} \tag{23}$$

In formula (23), ϖ_{1Q} and ϖ_{2Q} represent the learnable weights for source score calculation, and V_{N_e} represents the node set.

To ensure scalability, subgraph sampling is used for each event during training. A node's neighbors are sampled with an upper bound on the probability of sampling to form a mini-batch subgraph. Edge features and time encodings are normalized within the batch. For large-scale cross-event graphs, event aggregation (event local subgraphs) is prioritized before batch-level inference to avoid cross-event information leakage.

This paper aggregates multi-hop propagation information using a multi-layer T-GAT. For interpretability, the attention weight matrix and edge credibility of each layer are saved, and high-weight paths are output as candidate propagation paths according to weighted summation and ranking.

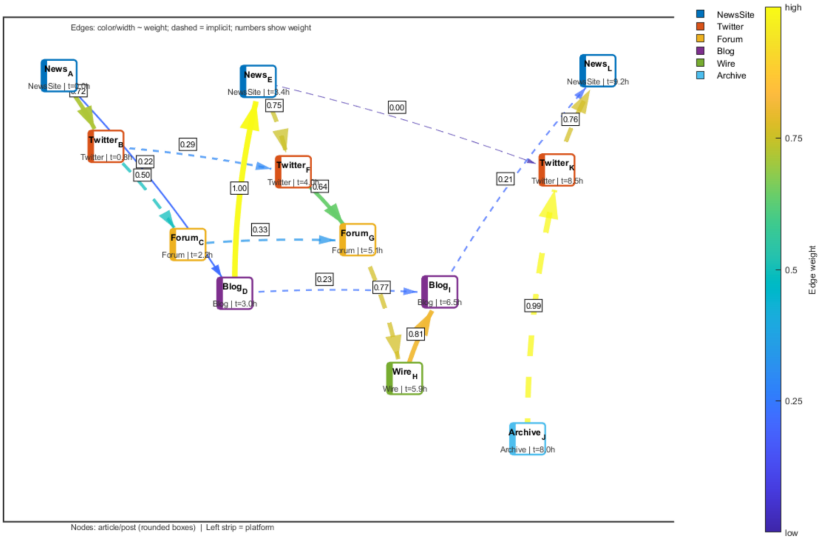

FIGURE 2. Constructed propagation map

GAT integrates dual temporal features, converting absolute event publication timestamps into learnable relative temporal embeddings and capturing long-term periodic patterns through sine-cosine positional encoding. It directly calculates the precise time difference (in minutes) of propagation between nodes and quantifies the time decay effect using an exponential decay function. These two temporal representations are concatenated and input into the graph attention layer, dynamically balancing semantic similarity and temporal proximity in edge weight calculation to ensure the model learns temporal patterns while preserving precise temporal relationships.

In this framework, "multi-source edge information" refers to the diverse relationship types extracted from heterogeneous propagation channels. It integrates three types of edge sources: first, explicit propagation relationships native to the platform, including forwarding, citation, reply, and hyperlink citation; second, implicit association edges calculated based on multimodal content similarity, automatically established when the similarity of textual semantics, visual content, or audio features of posts from different platforms exceeds a threshold; and third, potential propagation chains inferred through temporal proximity, capturing cross-platform content diffusion without explicit citations but with closely related publication times. Each edge type is assigned a one-hot encoded identifier and combined with a corresponding relationship strength weight, enabling T-GAT to distinguish the contribution of different propagation mechanisms to source tracing, such as prioritizing direct forwarding paths rather than coincidental propagation based solely on content similarity.

Although the main experimental verification in this paper is based on the scenario of news event tracing, the proposed multimodal Transformer–T-GAT hybrid framework also has clear applicability in event tracing tasks in complex industrial systems. This type of data is highly similar to the multimodal event propagation in social media in terms of modal structure, temporal scale, and propagation relationship, all exhibiting the characteristics of "multimodal observation + temporal association + propagation/dependency structure". In this application scenario, the multimodal encoding and two-stage cross-modal attention mechanism in this paper can be used to perform unified representation learning on manufacturing data from different sources. The visual encoder corresponds to the fabric defect image, the text encoder models the logs and process records, and the temporal embedding describes the evolution of process and equipment status. Furthermore, the temporal perception graph structure modeled by T-GAT can be naturally mapped to a process dependency graph or an equipment-batch association graph to characterize the propagation and accumulation mechanism of defects in the production chain. Therefore, the task of locating the source of an event can be transformed into the problem of identifying the root cause of the defect, while the path reconstruction results provide an interpretable basis for quality traceability and responsibility location.

N. LOSS FUNCTION DESIGN AND JOINT TRAINING OPTIMIZATION

The multimodal contrast loss uses InfoNCE (Information Noise-Contrastive Estimation) contrast loss, as shown in (24).

$$\mathrm{Loss}_{\mathrm{NCE}}^{\mathrm{text,img}} = -\frac{1}{|\mathrm{batch}|}\sum_{i\in\mathrm{batch}}\log\frac{\exp(\mathrm{sim}(\bar{h}_i^{\mathrm{text}},\bar{h}_i^{\mathrm{img}})/\varsigma)}{\sum_{j\in B}\exp(\mathrm{sim}(\bar{h}_i^{\mathrm{text}},\bar{h}_j^{\mathrm{img}})/\varsigma)} \tag{24}$$

In formula (24), ς represents the temperature parameter. Importantly, negative sample construction strictly follows these principles: B represents the current training batch, but negative samples only include samples from different events within the same batch; multimodal samples within the same event are treated as positive sample pairs and will

not be used as negative samples of each other; training data is strictly divided by event to ensure that all samples for each event are either entirely in the training set or entirely in the validation/test set, completely avoiding label leakage.

This paper employs the InfoNCE contrastive loss for multimodal contrastive loss, where negative sample construction comprises two key components: (1) samples from other modalities within the same event (e.g., image samples in text-image pairs are used as negative samples for text) to enhance intramodal consistency; and (2) samples from different events are used as cross-event negative samples to improve event discrimination. This design enables the model to simultaneously learn intermodal alignment relationships and inter-event discriminative characteristics, effectively avoiding the representation collapse problem that may result from using only intra-event negative samples.

Similarly, it defines the other two modalities: text and audio, and image and audio. Finally, it performs a weighted sum to obtain the overall contrast loss Losscontra.

The downstream classification loss (text classification and sentiment recognition) is expressed as (25).

$$\begin{cases} \text{Loss}_{\text{class}} = \dfrac{1}{|\text{batch}|} \displaystyle\sum_{i \in \text{batch}} \text{CE}(\hat{y}_{\text{class}}, y_i^{\text{class}}), \\ \text{Loss}_{\text{sent}} = \dfrac{1}{|\text{batch}|} \displaystyle\sum_{i \in \text{batch}} \text{CE}(\hat{y}_{\text{sent}}, y_i^{\text{sent}}). \end{cases} \tag{25}$$

In formula (25), y_i^{class} and y_i^{sent} represent the one-hot vector and sentiment label of the text classification truth, respectively, and CE represents the cross entropy.

To directly optimize the path reconstruction quality, a Soft-Jaccard loss is applied for the predicted edge weights, as shown in formula (26).

$$\text{Loss}_{\text{IoU}}(e) = 1 - \frac{\sum_{(i,j) \in E_{S_e}} \hat{\rho}_{ij} \phi_{ij}}{\sum_{(i,j) \in E_{S_e}} \hat{\rho}_{ij} + \sum_{(i,j) \in E_{S_e}} \phi_{ij} - \sum_{(i,j) \in E_{S_e}} \hat{\rho}_{ij} \phi_{ij} + \epsilon} \tag{26}$$

In formula (26), ϕ_{ij} represents the true binary vector, and $\hat{\rho}_{ij}$ represents the predicted vector. The overall edge weight loss is

$$L_{path} = \frac{1}{|E_{batch}|} \sum_{e \in E_{batch}} Loss_{IoU}(e)$$

The true binary vectors are obtained by combining manual annotation and system parsing. For propagation paths with clear forwarding/referencing relationships, the metadata provided by the platform is used directly. For propagation chains that lack explicit connections but have highly similar content, domain experts are used to annotate them based on time sequence and content similarity. At the same time, the event source information verified by fact-checking agencies is used as an authoritative reference to ensure the accuracy and reliability of the tags.

For source sorting, event-level cross entropy is used. An unnormalized score is output for each event node, and the probability is obtained through softmax. The cross entropy is then applied to the true source, as shown in formula (27).

$$\text{Loss}_{\text{src}} = -\frac{1}{|E_{\text{batch}}|} \sum_{e \in E_{\text{batch}}} \sum_{i \in V_{N_e}} y_i^{\text{src}} \log \pi_i \tag{27}$$

In formula (27), y_i^{src} represents the true source. The overall loss is the weighted sum of several items, as shown in formula (28).

$$\begin{aligned} \text{Loss}_{\text{total}} = {} & \chi_{\text{contra}} \text{Loss}_{\text{contra}} + \chi_{\text{class}} \text{Loss}_{\text{class}} + \chi_{\text{sent}} \text{Loss}_{\text{sent}} \\ & + \chi_{\text{path}} \text{Loss}_{\text{path}} + \chi_{\text{src}} \text{Loss}_{\text{src}} \end{aligned} \tag{28}$$

In formula (28), χ_{contra}, χ_{class}, χ_{sent}, χ_{path}, and χ_{src} represent the weights corresponding to different losses. The weights of each loss term were determined through a grid search on the validation set, rather than using a learnable weighting mechanism. This paper systematically tested the impact of different weight combinations on the main metrics (Top-1 source accuracy and Path-IoU) on the validation set, and ultimately selected the weight configuration that achieves the optimal balance between the two metrics: $\chi_{\text{contra}} = 0.3$, $\chi_{\text{class}} = 0.2$, $\chi_{\text{sent}} = 0.1$, $\chi_{\text{path}} = 0.25$, $\chi_{\text{src}} = 0.15$. This tuning strategy ensures that the losses from each task contribute reasonably to the training process, avoiding the problem of certain loss terms dominating the optimization direction due to excessively large differences in magnitude.

III. EXPERIMENTAL SETUP AND EVALUATION PLAN

A. EXPERIMENTAL DATA

The experimental data is organically integrated from three sources: publicly available multimodal and news-related datasets (FakeNewsNet, Fakeddit, and Twitter/Weibo) as the foundational annotation sources; fact-checking websites (PolitiFact, Snopes) and mainstream news organizations (Reuters, BBC (British Broadcasting Corporation)) crawled from verified events to obtain clear source labels and authoritative text; and cross-platform communication samples crawled from social media and video platforms (Twitter/X, Weibo, YouTube) to expand multimodal evidence and transmission chain information.

The final corpus covers 5,000 independent news events, approximately 30,000 news posts/article samples (including body text), approximately 25,000 associated images, and approximately 6,000 video/audio clips.

Complete metadata is preserved for each sample: publication timestamp, source URL, publishing account/media ID, retweet/citation relationships, and language tags. Approximately 80% of the samples achieved modality alignment at the event level (each event contains at least text and images or text and audio). All events are manually or systematically annotated with the "earliest trusted source" for provenance assessment. The dataset is segmented by event time, with a training/validation/testing ratio of 70%/15%/15% per event

to prevent future information leakage and ensure generalization across time.

B. EXPERIMENTAL PLAN AND PARAMETER SETTINGS

The comparison models include (1) Multimodal Transformer + T-GAT (the method in this paper); (2)

Multimodal Transformer + GAT; (3) Multimodal Transformer + GCN (Graph Convolutional Networks); (4)

VisualBERT-GAT; (5) LXMERT (Learning Cross-Modality Encoder Representations from Transformers)-GAT;

(6) BiLSTM-GAT; (7) T-GAT (using only text features); (8) GraphSAGE (Graph Sample and Aggregate) (using multimodal splicing features for nodes). For each model, the Top-K accuracy, MRR (Mean Reciprocal Rank), Edge-F1, Path-IoU, AUC (Area Under Curve) and other indicators are recorded, and the mean and confidence interval are obtained after 5 training runs with different random seeds. At the same time, ablation experiments (removing image/audio/text) are performed to quantify the contribution of each module. All models maintain the same training schedule and approximate hyperparameter search range.

The validation set is used for hyperparameter selection and finally a rigorous comparison is made on the test set.

In robustness experiments, the test set is subjected to various scenarios, including random noise, key node deletion, the presence of spurious edges, and edge weight perturbations. The performance degradation of each model is evaluated to verify the robustness of the reconstruction. Efficiency evaluations measured each model's training time, single-event inference latency, and graphics memory usage on the same hardware to assess engineering deployability. Finally, the generated propagation paths are output and visualized using attention to verify the legitimacy of the evidence chain for high-confidence paths.

Table 1 shows the experimental parameter settings.

TABLE 1. Experimental parameter settings

Parameters	Value	Parameters	Value
Hidden dimensions	768	Number of multi-head attention heads	8
Learning rate	2×10^{-5}	Number of fusion layers	6
Graph network learning rate	1×10^{-4}	Number of T-GAT layers	4
Batch size (sample level)	32	Comparative learning temperature	0.07
Number of graph batches/subgraph nodes	512	Comparative negative sample queues	4096
Number of epochs	20	Sorting margin	0.1
Early-stop tolerance	5 epochs	Time decay coefficient	Initial 1×10^{-3}
Dropout	0.1	Gradient clipping	1

IV. ANALYSIS OF NEWS EVENT TRACING RESULTS

A. TRACEABILITY PERFORMANCE EVALUATION

The traceability performance evaluation results are shown in Figure 3.

Figure 3 shows that Multimodal Transformer + T-GAT leads in all provenance metrics, with a Top-1 accuracy of 0.92, a 0.06 improvement over Multimodal Transformer + GAT and a 0.09 improvement over VisualBERT + GAT. Top-3 and Top-5 accuracy are 0.94 and 0.98, respectively. In terms of ranking quality, Multimodal Transformer + T-GAT achieves an MRR of 0.88, higher than Multimodal Transformer + GAT's 0.82 (an improvement of 0.06), indicating that the proposed model achieves better overall ranking. Multimodal Transformer + T-GAT achieves an AUC of 0.95, surpassing Multimodal Transformer + GAT's 0.92 (an improvement of 0.03), demonstrating stronger discriminative capabilities. All methods with multimodal Transformer features and an attention-based graph network (multimodal Transformer + T-GAT/GAT) significantly outperform baselines using only text or traditional graph convolution in both Top-k and MRR (BiLSTM + GAT Top-1 accuracy of 0.65 and MRR of 0.60; T-GAT Top-1 accuracy of 0.72). GraphSAGE and multimodal Transformer + GCN performed in the middle, with Top-1 accuracy of 0.78 and 0.80, respectively.

This shows that the quality of modal representation and graph reasoning mechanisms has a significant impact on the final traceability.

B. VERIFICATION OF ROBUST PERFORMANCE OF PROPAGATION PATH RECONSTRUCTION

To explore the robustness of propagation path reconstruction, it randomly inserts noise nodes into the graph. It connects them, deletes key intermediate/early nodes in the propagation chain, inserts misleading high-similarity but non-propagation edges into the graph, and randomly perturbs/noises the existing edge weights in four situations to explore its robustness. The results are shown in Figure 4.

In the experiments with random noise nodes and edge weight perturbations shown in Figure 4, the Multimodal Transformer+T-GAT performs best. With random noise nodes, the Path-IoU is 0.7, the Edge-Accuracy is 0.93, and the Edge-F1 is 0.88, which are 0.05/0.02/0.04 higher than the Multimodal Transformer+GAT, respectively. With edge weight perturbations, the Path-IoU is 0.66, the Edge-Accuracy is 0.92, and the Edge-F1 is 0.86, also better than the 0.62/0.90/0.82 of the Multimodal Transformer+GAT.

These metrics show that when noise nodes are present or existing edge weights are randomly perturbed, T-GAT's temporal encoding and edge-level attention mechanism effectively suppress the impact of error propagation on path reconstruction. Multimodal features enhance the semantic connections between nodes, minimizing the impact of edge weight perturbations on the overall path. Other models such as BiLSTM-GAT and T-GAT alone have Path-IoUs of 0.44/0.53 under random noise nodes and 0.40/0.52 under edge weight perturbations, respectively. This shows that

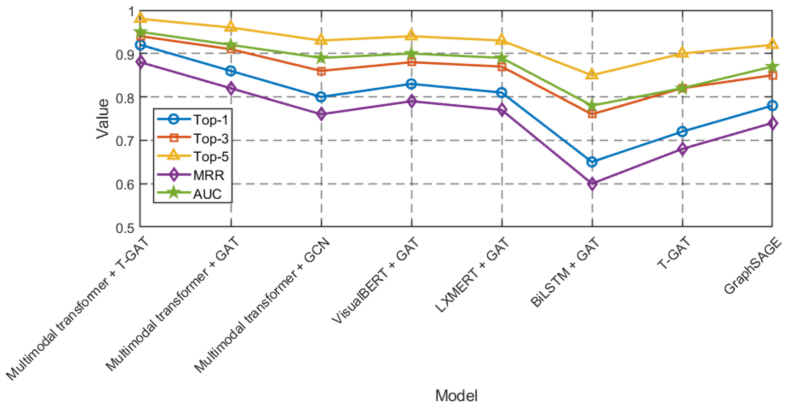

FIGURE 3. Traceability performance evaluation

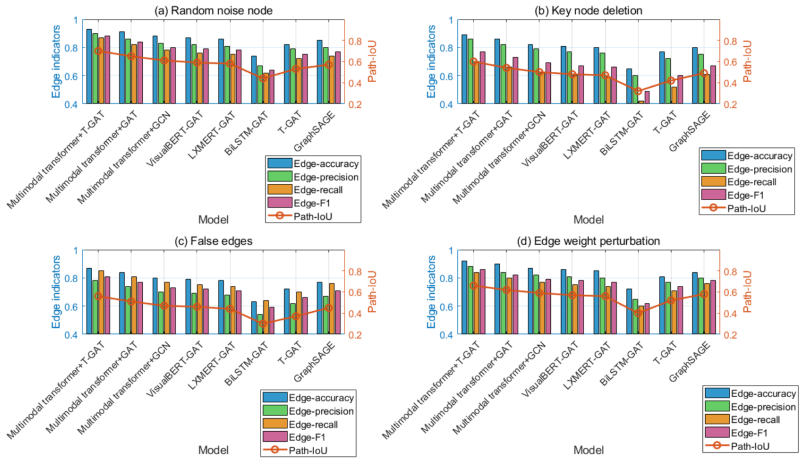

FIGURE 4. Robust performance verification results of propagation path reconstruction. Figure 4 (a) Random noise nodes; Figure 4 (b) Key node deletion; Figure 4 (c) False edges; Figure 4 (d) Edge weight perturbation.

single-modal or weak node features are easily misled in noisy environments and their robustness is obviously insufficient.

In experiments with key node removal and the presence of spurious edges, Multimodal Transformer+T-GAT maintains its lead. With key node removal, its Path-IoU reaches 0.6 and its Edge-F1 reaches 0.77, significantly higher than Multimodal Transformer+GAT (Path-IoU 0.54, Edge-F1 0.73) and BiLSTM-GAT (Path-IoU 0.32, Edge-F1 0.49). With spurious edges present, Multimodal Transformer+T-GAT achieves Path-IoU 0.56 and Edge-F1 0.81, still significantly outperforming Multimodal Transformer+GAT (0.51/0.77) and BiLSTM-GAT (0.30/0.59). This shows that when the core nodes of the propagation chain are missing or misleading edges with high similarity are artificially added to the graph, T-GAT can effectively identify the real propagation link through temporal attention and multimodal semantic consistency, ensuring the relative integrity and accuracy of path reconstruction. Other models are prone to path breakage or misconnection when key nodes are missing or false edges are encountered, resulting in a greater decrease in Path-IoU and Edge-F1.

C. TEXT CLASSIFICATION AND SENTIMENT RECOGNITION PERFORMANCE

This paper analyzes the quantitative results of the candidate models for the three tasks of text classification (factualness and stance) and sentiment recognition. The results are shown in Figure 5.

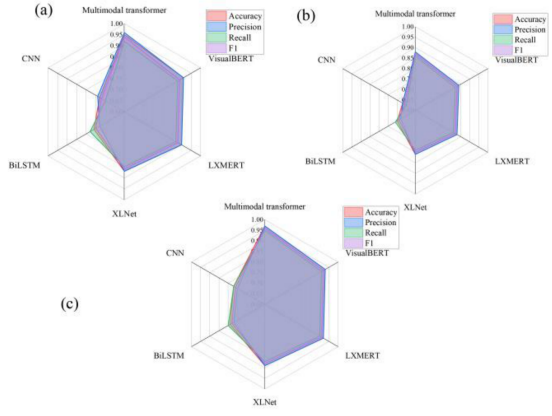

FIGURE 5. Text classification and sentiment recognition performance. Figure 5 (a) Factualness; Figure 5 (b) Stance; Figure 5 (c) Sentiment

As shown in Figure 5, the Multimodal Transformer performs well across all three tasks, achieving the highest performance in sentiment recognition, with accuracy of 0.97, precision of 0.97, recall of 0.95, and F1 of 0.96. In factual classification, it achieves accuracy of 0.95, precision of 0.96, recall of 0.92, and F1 of 0.94. However, its performance in stance classification is relatively poor, with accuracy of 0.88, precision of 0.88, recall of 0.86, and F1 of 0.87. Compared to other models, VisualBERT, LXMERT, and XLNet (EXtreme Language Modeling) all achieve accuracy of 0.89 or higher in sentiment recognition, while BiLSTM and CNN perform relatively poorly in all tasks, achieving accuracy of 0.7 and 0.67, respectively, in stance classification. This demonstrates that deep multimodal features and the attention mechanism significantly improve performance on these tasks. The differences between the same tasks also reflect that multimodal information has a stronger ability to extract emotional and factual features, while stance classification is more dependent on text context and implicit opinions, which cannot be fully captured by simple feature fusion.

D. MODAL CONTRIBUTION ANALYSIS

To explore the contribution of each modality, the experiment uses an ablation method, removing each modality in turn and examining its impact. First, key traceability metrics (Top-3 accuracy, Path-IoU, and Edge-accuracy) are presented for different modality combinations. Next, the contribution of each modality to performance is analyzed. Figure 6 shows the results of the modal contribution analysis.

In Figure 6(b), "contribution" is defined as the relative contribution of each modality to the improvement of system performance. It is obtained by calculating the performance difference between the complete multimodal system (text + image + audio) and the system after removing a single modality, and then normalizing the result. Contribution = (Complete system performance - System performance after removing the modality) / Σ(Complete system performance - System performance after removing all modalities) $\times$ 100%. The performance metrics comprehensively consider three dimensions: Top-3 accuracy, path IoU, and edge accuracy, ensuring the comprehensiveness and robustness of the contribution assessment.

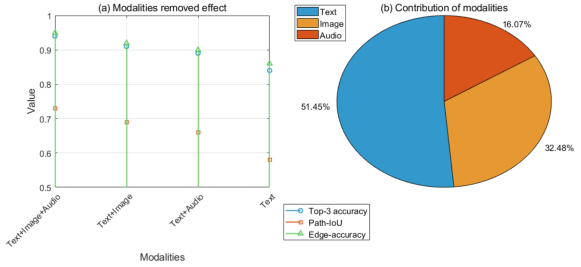

FIGURE 6. Modal contribution analysis results. Figure 6 (a) Effect after modal removal; Figure 6 (b) Contribution level

Figure 6 shows that the text modality contributes most to news event attribution performance. When used independently, it achieves a Top-3 accuracy of 0.84, a Path-IoU of 0.58, and an Edge-accuracy of 0.86, contributing 51.45% to the overall multimodal fusion. After removing the audio modality (text + image), the Top-3 accuracy improves to 0.91, the Path-IoU of 0.69, and the Edge-accuracy of 0.92, demonstrating that the image modality effectively supplements text information, contributing 32.48%. When removing the image modality (text + audio), the Top-3 accuracy is 0.89, the Path-IoU of 0.66, and the Edge-accuracy of 0.90, indicating that audio's contribution to attribution performance is limited, accounting for only 16.07%.

Comprehensive analysis shows that the combined use of three modalities (text + image + audio) can achieve optimal performance with Top-3 accuracy of 0.94, Path-IoU of 0.73, and Edge-accuracy of 0.95, demonstrating that the modalities complement each other. Text provides core semantic information, images enhance evidence verification capabilities, and audio provides auxiliary emotional and contextual clues, significantly improving the overall performance of the multimodal Transformer in news event tracing.

E. ATTENTION VISUALIZATION

The attention visualization diagram is shown in Figure 7.

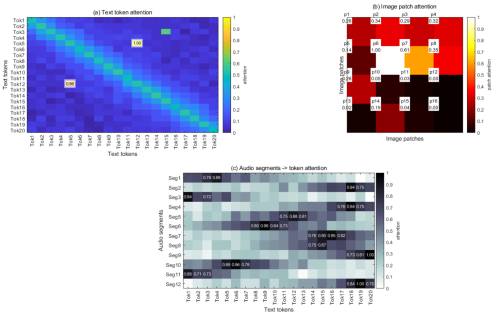

FIGURE 7. Attention Visualization. Figure 7 (a) Text token attention heatmap; Figure 7 (b) Image patch attention heatmap; Figure 7 (c) Cross-modal attention heatmap for audio and text

In Figure 7(a), the text token attention heatmap shows the dependencies between different tokens.

Brighter areas correspond to high attention values, indicating that the model assigns higher weights to key tokens in the sentence (such as the 5th, 12th, and 15th tokens), reflecting its focus on semantic associations within the text. The image patch attention heatmap in Figure 7(b) reveals the model's attention to different regions of the image. Specific patches (such as the 6th and 7th patches) are highlighted, demonstrating that the model is able to identify areas that are more important for the task in visual information processing. The audio and text attention heatmap in Figure 7(c) shows the interaction between audio time segments and text tokens. Darker areas indicate that the model is able to effectively align key audio information with relevant text information, demonstrating its ability to integrate multimodal information. Taken together, these three attention

heatmaps demonstrate that the model is able to highlight key information across different modalities, and its high attention values reflect its sensitivity to input features. The attention patterns of text, images, and audio complement each other, helping to explain the decision-making basis of the model when handling complex multimodal tasks, while providing an intuitive reference for further optimizing the feature representation of specific modalities or enhancing cross-modal associations.

F. TEMPORAL DYNAMIC MODELING ANALYSIS

To investigate the impact of temporal dynamic modeling on traceability and path reconstruction, it performs a performance analysis based on different propagation time periods. The results of this temporal dynamic modeling analysis are shown in Figure 8.

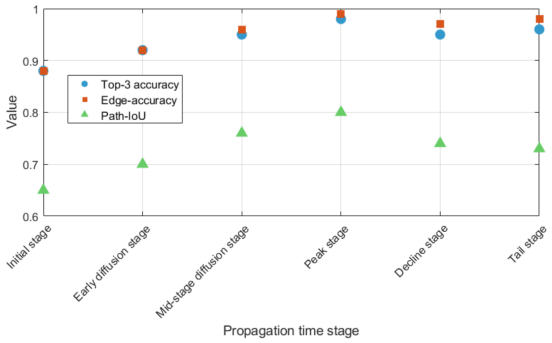

FIGURE 8. Time dynamic modeling analysis results

Figure 8 shows that temporal dynamics modeling significantly improves the performance of news event tracing and path reconstruction. In the initial phase, the Top-3 accuracy is 0.88, the Edge-accuracy is 0.88, and the Path-IoU is 0.65, indicating that limited information in the early stages made tracing and path reconstruction difficult. As the event enters its early diffusion phase, these metrics improve to 0.92 Top-3 accuracy, 0.92 Edge-accuracy, and 0.70 Path-IoU, demonstrating that the increasing abundance of time series information facilitates model inference. During the mid-stage of diffusion, the Top-3 accuracy reaches 0.95, the Edge-accuracy reaches 0.96, and the Path-IoU reaches 0.76. Performance is optimal during the peak phase, with Top-3 accuracy of 0.98, Edge-accuracy of 0.99, and 0.80 Path-IoU, demonstrating that the model fully utilizes information from time periods with high transmission volume for path reconstruction. In the decay and tail stages, the Top-3 accuracy is 0.95 and 0.96, the Edge-accuracy is 0.97 and 0.98, and the Path-IoU is 0.74 and 0.73, respectively. This shows that although the propagation is weakened, the temporal dynamic modeling can still maintain a high traceability performance. The overall trend is to first increase and then stabilize with the propagation process, reflecting the effective use of time series information by the model.

G. RESOURCE AND EFFICIENCY ANALYSIS

The experiment measures resource and efficiency results in a unified experimental environment. Training time is the total GPU (graphics processing unit) hours required for complete training to convergence, single-event inference latency is the average number of milliseconds required for end-to-end inference of a single sample, model parameter count is in millions (M), GPU memory usage is peak usage (GB), and energy consumption is the average energy consumption per inference (J). The resource and efficiency analysis results are shown in Table 2.

The experiment was conducted in a unified hardware environment: a server equipped with four NVIDIA A100 80GB GPUs, an Intel Xeon Platinum 8380 CPU (2.3GHz, 40 cores), and 1TB of RAM. Energy consumption data was collected in real-time using the NVIDIA-SMI tool, combined with CPU power consumption monitoring via the Intel RAPL interface. The total energy consumption for a single inference run was calculated by integration, excluding data loading and preprocessing overhead. Single-event inference latency and energy consumption measurements were performed with a batch size of 1 to reflect single-sample processing performance in real-world application scenarios; training time metrics were measured based on a batch size of 32 as set in Table 1. GPU memory usage was recorded as peak memory usage during training. All measurements are averages of five independent runs, with a confidence interval of $\pm 2\%$.

TABLE 2. Resource and efficiency analysis results

Model	Training time (h)	Single-event inference latency (ms)	Model parameter count (M)	GPU memory usage (GB)	Energy consumption (J/inference)
Multimodal Transformer + T-GAT	98.60	420.77	315.22	25.40	72.15
Multimodal Transformer + GAT	128.45	512.37	412.58	31.24	95.27
Multimodal Transformer + GCN	114.30	482.12	389.41	29.10	88.66
VisualBERT + GAT	106.75	460.05	387.05	27.85	84.33
LXMERT + GAT	103.25	435.90	328.49	26.05	75.82
BiLSTM + GAT	36.55	210.12	12.78	4.95	15.30
T-GAT	52.80	230.44	25.34	6.20	18.90
GraphSAGE	60.10	250.67	45.67	8.40	22.45

In Table 2, the multimodal Transformer + T-GAT has a training time of 98.60 hours, a single-event inference latency of 420.77 ms, a model parameter count of 315.22M, a GPU memory usage of 25.40 GB, and an energy consumption of 72.15 J/inference. Compared with other multimodal combinations such as the multimodal Transformer + GAT (training time of 128.45 hours, inference latency of 512.37 ms, a parameter count of 412.58M, a GPU memory usage of 31.24 GB, and an energy consumption of 95.27 J/inference), it shows better training and inference efficiency while reducing memory and energy consumption. Compared to lightweight models such as BiLSTM + GAT (training time 36.55 hours, inference latency 210.12 milliseconds, parameter count 12.78MB, GPU memory 4.95 GB, energy consumption 15.30 J/inference) and T-GAT (training time 52.80 hours, latency 230.44 milliseconds, parameter count 25.34MB, GPU memory 6.20 GB, energy consumption 18.90 J/inference), the combination of the Multimodal Transformer and T-GAT, while more resource-intensive, demonstrates reasonable efficiency and affordable computational overhead in a trade-off between improved accuracy and traceability performance. Other multimodal Transformer combinations, such as Multimodal Transformer + GCN, VisualBERT + GAT, and LXMERT + GAT, all indicate that the proposed multimodal fusion model can reduce training and inference costs while maintaining performance.

V. CONCLUSION

This paper constructs a hybrid framework that fuses a multimodal Transformer with T-GAT, achieving cross-modal alignment by encoding features for text, image, and audio modalities separately and incorporating temporal and platform information. Joint representation learning is then performed using the multimodal Transformer, combined with T-GAT for temporal propagation reasoning, enabling event source location and propagation path reconstruction.

Experimental results demonstrate that this method achieves a Top-1 provenance accuracy of 0.92, a 0.09 improvement over VisualBERT+GAT, and Top-3/Top-5 accuracy of 0.94/0.98, respectively. The Path-IoU remains at 0.66 even under edge weight perturbations, demonstrating strong robustness. However, this paper still has some shortcomings. The audio and video modality features are limited by the expressive power of the pre-trained model, limiting its adaptability to low-quality data. Furthermore, the propagation graph modeling relies primarily on explicit side information, requiring further optimization to capture potential implicit relationships. Future work can consider applying more powerful large models and causal graph modeling methods to enhance cross-modal evidence fusion and interpretability.

AUTHOR CONTRIBUTIONS

Conceptualization –Yao Li, Shaohua Wei, Tian Bao and Wei Gao; methodology – Yao Li, Shaohua Wei, Tian Bao and Wei Gao; investigation – Yao Li, Shaohua Wei and Tian Bao; writing-original draft preparation – Yao Li, Shaohua Wei, Tian Bao and Wei Gao. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, Enterprise service project "Construction of China Digital Sports City Influence Evaluation System" (No.:
2023004).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Ren, H. Dong, A. Popovic, G. Sabnis, and J. Nickerson, "Digital platforms in the news industry: how social media platforms impact traditional media news viewership," European journal of information systems, vol. 33, no. 1, pp. 1-18, 2024, doi: 10.1080/0960085X.2022.2103046.

[2] A. Denisova, "Viral journalism," Strategy, tactics and limitations of the fast spread of content on social media: Case study of the United Kingdom quality publications. Journalism, vol. 24, no. 9, pp. 1919-1937, 2023, doi: 10.1177/14648849221077749.

[3] K. Kaur and S. Gupta, "Towards dissemination, detection and combating misinformation on social media: a literature review," Journal of business & industrial marketing, vol. 38, no. 8, pp. 1656-1674, 2023, doi: 10.1108/JBIM-02-2022-0066.

[4] A. Jarrahi and L. Safari, "Evaluating the effectiveness of publishers' features in fake news detection on so-cial media," Multimedia Tools and Applications, vol. 82, no. 2, pp. 2913-2939, 2023, doi: 10.1007/s11042-022-12668-8.

[5] S. Biradar, S. Saumya, and A. Chauhan, "Combating the infodemic: COVID-19 induced fake news recogni-tion in social media networks," Complex & Intelligent Systems, vol. 9, no. 3, pp. 2879-2891, 2023, doi: 10.1007/s40747-022-00672-2.

[6] T. Aharoni, K. Tenenboim-Weinblatt, N. Kligler-Vilenchik, P. Boczkowski, K. Hayashi, E. Mitchelstein, et al., "Trust-oriented affordances: A five-country study of news trustworthiness and its socio-technical articulations," New Media & Society, vol. 26, no. 6, pp. 3088-3106, 2024, doi: 10.1177/14614448221096334.

[7] S. K. Uppada and P. B. S. Patel, "An image and text-based multimodal model for detecting fake news in OSN's," Journal of Intelligent Information Systems, vol. 61, no. 2, pp. 367-393, 2023, doi: 10.1007/s10844-022-00764-y.

[8] B. Palani, S. Elango, and K. V. Viswanathan, "CB-Fake: A multimodal deep learning framework for auto-matic fake news detection using capsule neural network and BERT," Multimedia Tools and Appli-cations, vol. 81, no. 4, pp. 5587-5620, 2022, doi: 10.1007/s11042-021-11782-3.

[9] V. Kishore and M. Kumar, "Enhanced multimodal fake news detection with optimal feature fusion and modified bi-lstm architecture," Cybernetics and systems, vol. 56, no. 6, pp. 684-714, 2025, doi: 10.1080/01969722.2023.2175155.

[10] Y. Guo, "A mutual attention based multimodal fusion for fake news detection on social network," Ap-plied Intelligence, vol. 53, no. 12, pp. 15311-15320, 2023, doi: 10.1007/s10489-022-04266-w.

[11] R. Kumari, N. Ashok, P. K. Agrawal, T. Ghosal, and A. Ekbal, "Identifying multimodal misinformation leveraging novelty detection and emotion recognition," Journal of Intelligent Information Systems, vol. 61, no. 3, pp. 673-694, 2023, doi: 10.1007/s10844-023-00789-x.

[12] G. Zhang, A. Giachanou, and P. Rosso, "SceneFND: Multimodal fake news detection by modelling scene context information," Journal of Information Science, vol. 50, no. 2, pp. 355-367, 2024, doi: 10.1177/01655515221087683.

[13] M. Bhadri, "BiLSTMs and BPE for English to Telugu CLIR," J. Electrical Systems, vol. 20, no. 3, pp. 2022-2029, 2024, doi: 10.52783/jes.1798.

[14] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," Journal of healthcare engineering, vol. 2022, no. 1, pp. 1-17, 2022, doi: 10.1155/2022/3498123.

[15] M. A. Shah, M. J. Iqbal, N. Noreen, and I. Ahmed, "An automated text document classification framework using BERT," Int. J. Adv. Comput. Sci. Appl, vol. 14, no. 3, pp. 279-285, 2023, doi: 10.14569/IJACSA.2023.0140332.

[16] M. I. Salih, S. M. Mohammed, A. K. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-Tuning BERT for Auto-mated News Classification," Engineering, Technology & Applied Science Research, vol. 15, no. 3, pp. 22953-22959, 2025, doi: 10.48084/etasr.10625.

[17] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake News Classification using transformer based enhanced LSTM and BERT," International Journal of Cognitive Computing in Engineering, vol. 3, no. 1, pp. 98-105, 2022, doi: 10.1016/j.ijcce.2022.03.003.

[18] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT against traditional machine learning models in text classification," Journal of Computational and Cognitive Engineering, vol. 2, no. 4, pp. 352-356, 2023, doi: 10.47852/bonviewJCCE3202838.

[19] M. Visweswaran, J. Mohan, S. S. Kumar, and K. P. Soman, "Synergistic detection of multimodal fake news leveraging TextGCN and Vision Transformer," Procedia Computer Science, vol. 235, no. 1, pp. 142-151, 2024, doi: 10.1016/j.procs.2024.04.017.

[20] M. Kunesova, Z. Zajic, L. Smidl, and M. Karafiat, "Comparison of wav2vec 2.0 models on three speech processing tasks," International Journal of Speech Technology, vol. 27, no. 4, pp. 847-859, 2024, doi: 10.1007/s10772-024-10140-6.

[21] Z. Kozhirbayev, "Kazakh speech recognition: Wav2vec2," 0 vs. whisper. Journal of Advances in In-formation Technology, vol. 14, no. 6, pp. 1382-1389, 2023, doi: 10.12720/jait.14.6.1382-1389.

[22] M. Khan, P. N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: multimodal emotion recog-nition using cross-modal transformer-based feature fusion," Scientific reports, vol. 15, no. 1, pp. 5473-5485, 2025, doi: 10.1038/s41598-025-89202-x.

[23] B. Wang, Y. Feng, X. Xiong, Y. Wang, and B. Qiang, "Multi-modal transformer using two-level visual fea-tures for fake news detection," Applied Intelligence, vol. 53, no. 9, pp. 10429-10443, 2023, doi: 10.1007/s10489-022-04055-5.

[24] J. Wang, S. Qian, J. Hu, and R. Hong, "Positive unlabeled fake news detection via multi-modal masked transformer network," IEEE Transactions on Multimedia, vol. 26, no. 1, pp. 234-244, 2023, doi: 10.1109/TMM.2023.3263552.

[25] A. Khaliq, S. A. Awan, F. Ahmad, M. A. Zia, and M. Z. Iqbal, "Enhanced topic-aware summarization using statistical graph neural networks," Computers, Materials and Continua, vol. 80, no. 2, pp. 3221-3242, 2024, doi: 10.32604/cmc.2024.053488.