informatics

MDPI

Article

ETICD-Net: A Multimodal Fake News Detection Network via Emotion-Topic Injection and Consistency Modeling

Wenqian Shang [1,2], Jinru Yang [2], Linlin Zhang [2], Tong Yi [2,3,*] and Peng Liu [2,3,*]

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; shangwenqian@cuc.edu.cn
[2] School of Software, Xinjiang University, Urumqi 830000, China; 107552301697@stu.xju.edu.cn (J.Y.); zllnadasha@xju.edu.cn (L.Z.)
[3] The Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China
* Correspondence: yitong@mailbox.gxnu.edu.cn (T.Y.); liupeng@mailbox.gxnu.edu.cn (P.L.)

Abstract

The widespread dissemination of multimodal disinformation, which combines inflammatory text with manipulated images, poses a severe threat to society. Existing detection methods typically process textual and visual features in isolation or perform simple fusion, failing to capture the sophisticated semantic inconsistencies commonly found in false information. To address this, we propose a novel framework: Emotion-Topic Injection and Consistency Detection Network (ETICD-Net). First, a large language model (LLM) extracts structured sentiment and topic-guided signals from news texts to provide rich semantic clues. Second, unlike previous approaches, this guided signal is injected into the feature extraction processes of both modalities: it enhances text features from BERT and modulates image features from ResNet, thereby generating sentiment-topic-aware feature representations. Additionally, this paper introduces a hierarchical consistency fusion module that explicitly evaluates semantic coherence among these enhanced features. It employs cross-modal attention mechanisms, enabling text to query image regions relevant to its statements, and calculates explicit dissimilarity metrics to quantify inconsistencies. Extensive experiments on the Weibo and Twitter benchmark datasets demonstrate that ETICD-Net outperforms or matches state-of-the-art methods, achieving accuracy and F1 scores of 90.6% and 91.5%, respectively.

Keywords: emotion injection; fake news detection; multimodal; large language model

Academic Editor: Remo Pareschi

Received: 10 September 2025
Revised: 18 November 2025
Accepted: 21 November 2025
Published: 25 November 2025

Citation: Shang, W.; Yang, J.; Zhang, L.; Yi, T.; Liu, P. ETICD-Net: A Multimodal Fake News Detection Network via Emotion-Topic Injection and Consistency Modeling. *Informatics* **2025**, *12*, 129. https://doi.org/10.3390/informatics12040129

1. Introduction

In the digital information age, the rapid development of the Internet and social media has fundamentally transformed how news is disseminated and consumed. However, this convenience has also created fertile ground for the emergence and proliferation of a new type of threat—multimodal fake news [1]. Such content typically combines fabricated, highly inflammatory text with manipulated or semantically irrelevant images to mislead readers, manipulate public opinion, and even undermine social stability [2]. Figure 1 illustrates a multimodal fake news example sourced from social media. Compared to text-based misinformation, multimodal fake news poses greater detection challenges due to its heightened sensory impact and deceptive nature, emerging as one of the foremost challenges in fact-checking [3,4].

Text: Sharks in the mall!
After the hurricane sandy!

Text: New species of fish
found at Arkansas

Figure 1. Multimodal posts published on social media.

To address this challenge, researchers have proposed numerous deep learning-based multimodal fake news detection models. Early studies focused on extracting features separately from text and images, followed by simple fusion strategies such as concatenation, addition, or multiplication for integration and classification [5–7]. With the rise of attention mechanisms, many works began exploring interactions between modalities, such as utilizing cross-modal attention to capture fine-grained associations between text and images [8,9]. Recently, some studies have also attempted to leverage graph neural networks to model complex intermodal relationships, constructing image-text association graphs to capture deeper semantic interactions [10,11]. Furthermore, contrastive learning-based approaches have demonstrated potential by enhancing detection capabilities through learning representations of consistency and inconsistency across modalities [12,13]. Despite achieving specific successes, these methods still harbor limitations that constrain further performance improvements. For instance, existing feature fusion strategies remain largely superficial, relying on simple feature concatenation or early fusion while lacking deep modeling of complex, nonlinear interactions between modalities. This coarse-grained fusion approach struggles to capture subtle semantic inconsistencies. While some studies attempt to address this through multi-stage or hierarchical fusion, they fail to fully simulate the human process of integrating multimodal information [14]. Secondly, existing models lack high-level semantic guidance. Most existing work directly fuses low-level or mid-level features within data-driven frameworks, failing to incorporate high-level semantic priors like "emotion" or "topic" that humans use to identify misinformation. Fabricators of fake news often deliberately employ inflammatory emotions such as anger or fear, alongside trending topics, to mask false content and boost dissemination. Current models remain insensitive to such intentional semantic manipulation patterns. While some works have modeled consistency in shared latent spaces, they often lack explicit modeling within the context of specific, high-level semantic priors, such as emotion and topic, which are crucial for identifying sophisticated disinformation. Most approaches treat detection as an end-to-end black-box classification problem, failing to explicitly and structurally measure and quantify semantic consistency between textual descriptions and visual evidence—a core reasoning logic of human fact-checkers.

To address these limitations, this paper proposes a novel Emotion-Topic Injection and Consistency Detection Network (ETICD-Net). Its core hypothesis posits that "emotion" and "topic" can serve as highly efficient higher-order prior information to guide multimodal consistency reasoning. This design is grounded in the observation that fake news often deliberately employs specific emotional tones like exaggeration to amplify dissemination effects, while leveraging trending topics for disguise. ETICD-Net's core concept leverages large language models' (LLMs) robust zero-shot semantic understanding and generation capabilities to automatically extract high-level "emotion-topic" guidance signals from news texts. Subsequently, through an innovative hierarchical fusion framework, it explicitly evaluates the consistency of multimodal information within specific emotion-topic contexts. Specifically, by grounding the model's consistency assessment in these prior cues, it aims

to simulate human fact-checking processes—namely, "whether visual evidence aligns with the text's emotional and thematic claims". The model's interpretability is dual-faceted: it provides a clear discrepancy metric F_{diff} to quantify inconsistencies, while its decisions can be traced back to high-level semantic conflicts through case studies. The specific solution comprises two components: First, designing an emotion-theme injection mechanism that integrates LLM-generated structured guidance vectors into BERT text features and ResNet image features through gating and conditional instance normalization, yielding emotion-theme-aware multimodal representations. Second, constructing a hierarchical consistency fusion module that first simulates "text-image" interactions via cross-modal attention, then projects enhanced modal features into a common space to compute explicit differences, enabling refined, interpretable evaluation of multimodal semantic consistency. The core objective of this research lies not only in improving accuracy but also in learning quantifiable interpretability scores.

The main contributions of this paper can be summarized as follows:

1. Proposing a novel multimodal fake news detection framework (ETICD-Net): This framework integrates large language models as high-level semantic guides with traditional visual and linguistic encoders, introducing a new paradigm to the field.
2. Introduces an affect-theme injection method: This paper designs an injection scheme based on gating mechanisms and conditional instance normalization, successfully deep-fusing discrete affect-theme semantic information into continuous modal features to enhance feature discriminability.
3. Design of a hierarchical consistency fusion module: This module combines cross-modal attention with explicit divergence metrics to structurally evaluate consistency between image and text modalities, improving performance and model interpretability.

The remainder of this paper is structured as follows: Section 2 reviews relevant research work; Section 3 details the model architecture of ETICD-Net; Section 4 presents experimental results and analysis; Section 5 summarizes the entire paper and outlines future research directions. Section 6 discusses the limitations of this work and offers future directions.

2. Related Work

2.1. Application of Sentiment Analysis in Fake News Detection

Sentiment analysis aims to computationally identify, extract, and quantify subjective emotional information within text, emerging as a significant research domain in natural language processing [15–17]. In recent years, with the rapid advancement of deep learning technologies, models based on the Transformer architecture have become the mainstream approach for sentiment analysis. These models can more effectively capture contextual semantics and complex emotional expressions within text [18,19]. In fake news detection, sentiment features have garnered significant attention due to their strong indicative role. Research indicates that fake information often deliberately employs stronger, more extreme emotional vocabulary—such as anger or fear—to amplify its spread and incite reader emotions [20]. Based on this finding, numerous researchers have explored integrating sentiment analysis into detection frameworks. Early attempts primarily treated sentiment as an auxiliary feature. For instance, Bounaama et al. [21] enhanced BERT model performance through sentiment attention mechanisms, enabling the model to focus more on inflammatory sections within text. Kula et al. [22] utilized GloVe word embeddings with LSTM/GRU networks to capture sentiment semantics in sequential text. More complex models were proposed to explore the potential of sentiment interactions further.

However, most of the aforementioned methods still treat sentiment analysis as a supplementary feature, failing to fully leverage its potential in deep semantic understanding

and cross-modal interaction. To overcome this limitation, recent research has focused on constructing more complex sentiment interaction models. For instance, Zhang et al. [23] developed a graph attention network that integrates sentiment interaction, comparison, and heterogeneous knowledge to analyze sentiment semantic relationships from multiple dimensions. Within the multi-task learning framework, Kumari et al. [24] jointly accomplished fake news detection and sentiment classification by sharing representations, while Jiang et al. [25] enhanced detection performance by jointly training sentiment classifiers and stance detectors, leveraging their interdependencies. Furthermore, the role of sentiment signals in multimodal semantic alignment and contextual understanding is gaining increasing attention. Sun et al. strengthened semantic associations between text and images by introducing sentiment-enhanced cross-modal contrastive learning, employing attention mechanisms to achieve dynamic fusion of modal features [26]. Toughrai proposed an intermediate-layer sentiment representation method to capture subtle emotional patterns in text, enhancing the model's contextual understanding and constructing a more robust sentiment-assisted detection model [27]. Tan designed a sentiment-semantic interaction network, exploring the latent connections between content sentiment and comment semantics from both questioning and non-questioning perspectives, thereby deepening the understanding of user engagement mechanisms in fake news propagation [28].

Although these studies successfully demonstrated the effectiveness of emotional signals, most still treat emotion as an independent feature that needs to be added to the model. They fail to integrate emotion with the subject at a higher level or utilize it as a global guiding signal to dynamically modulate the multimodal fusion process, which is precisely one of the starting points of this work.

2.2. Multimodal Fake News Detection

Multimodal fake news detection aims to build more robust detection models by integrating information sources such as text and images. Methods based on cross-modal consistency modeling represent one of the mainstream paradigms. The core of such approaches lies in evaluating semantic consistency across modalities. Early works like the SAFE framework proposed by Zhou et al. [29] performed text-image matching through similarity metrics. Khattar et al.'s MVAE [6] employs variational inference to construct a shared latent space, addressing the issue of missing modalities. To enhance the precision of consistency metrics, Chen et al.'s CAFE framework [30] introduces KL divergence to quantify cross-modal ambiguity. At the same time, Sun et al.'s KDCN model incorporates external knowledge to simultaneously capture inconsistencies across modalities and knowledge levels [31]. Yu et al.'s SR-CIBN [32] dynamically balances consistency and inconsistency features through weight scores. Attention-based methods form another key branch, effectively capturing fine-grained interactions between modalities. Wu et al.'s MCAN [14] achieves deep fusion by stacking multiple co-attention layers, while L. Wang et al.'s COOLANT framework [33] employs contrastive learning to optimize alignment in the representation space. Recently, Y. Jiang et al.'s CMA model [34] leverages shared attention to enhance cross-modal features, demonstrating strong performance in few-shot scenarios. Other innovative architectures exist, such as Ying et al.'s BMR [35], which improves performance through multi-view feature representation and an enhanced expert network.

Despite significant progress in these approaches, they primarily rely on data-driven low-level feature interactions during the fusion process, lacking global guidance from high-level semantics such as sentiment and topic. Furthermore, consistency evaluation is often implicitly embedded in model design rather than explicitly modeled and quantified in a structured manner. The ETICD-Net model proposed in this paper addresses these challenges by introducing high-level semantic guidance and performing explicit consis-

tency evaluation. It is also important to distinguish our work from research in multimodal media forensics, which focuses on detecting low-level tampering traces (e.g., copy-move, compression artifacts) within images. While forensics is crucial for identifying technically manipulated media, ETICD-Net addresses a different challenge: semantic-level inconsistency between authentic-looking images and their accompanying text. These approaches are complementary in the fight against multimodal disinformation.

3. Method

3.1. Overall Architecture

The overall framework of the ETICD-Net model is shown in Figure 2, primarily consisting of four core components: (1) Sentiment Topic Guiding Module; (2) Feature Extraction and Sentiment Topic Injection Module; (3) Hierarchical Consistency Fusion Module; (4) Classifier. Given a multimodal news instance (x_t, x_v), where x_t represents text and x_v represents images, the model first utilizes a large language model (LLM) to extract a high-level semantic guidance vector $\mathbf{g}$ from x_t. Subsequently, $\mathbf{g}$ is injected into both BERT-based text features and ResNet-based image features to generate sentiment-theme-aware feature representations $\mathbf{T}_g$ and $\mathbf{V}_g$. Finally, these augmented features are fed into the hierarchical consistency fusion module, which captures intermodal inconsistencies by computing cross-modal attention and explicit differences, ultimately outputting the classification prediction result.

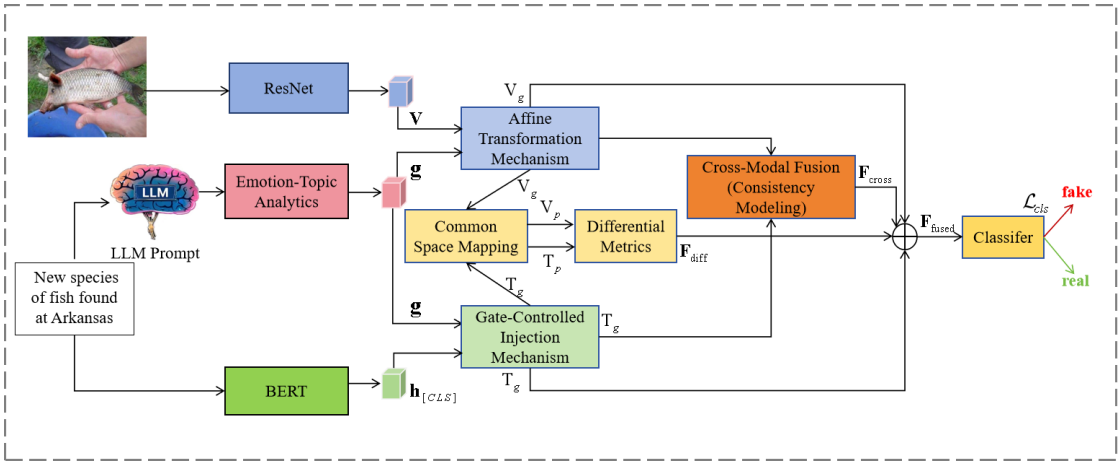

Figure 2. The overall framework of the ETICD-Net model, from left to right, consists of the sentiment topic guidance module, the feature extraction and sentiment topic injection module, the hierarchical consistency fusion module, and the classifier.

3.2. Feature Extraction and Sentiment-Theme Injection

This module aims to deeply integrate high-level emotional semantic information into the original features, providing a more semantically rich representation for subsequent consistency calculations.

3.2.1. Emotion-Themed Guided Generation

Given text x_t, this paper employs a carefully designed prompt template integrated with the ChatGLM's glm-4v-plus API from Zhipu AI (accessed on 20 September 2024) to generate structured descriptions. The prompt content is as follows: "We are conducting a fake news detection task. Below is the text content of the news article. Please help analyze the emotional and thematic information contained within this news item, outputting 1–3 key pieces of information in the specified format. Output format: Sentiment: [Sentiment1, Sentiment2, Sentiment3]. Topic: [Topic1, Topic2, Topic3]. Text: {x_t}". This design

ensures the large language model operates solely on the input text without incorporating external knowledge or metadata, thereby preserving the integrity of the closed-world assumption. These descriptions are parsed into a discrete sentiment label e and a sequence of topic keywords $w_1, w_2, \ldots, w_k$. This paper converts these discrete signals into dense vectors through embedding lookup and pooling operations. The sentiment label e undergoes one-hot encoding and is processed through a learnable embedding layer to yield the sentiment vector $\mathbf{e} \in \mathbb{R}^{d_e}$. Each topic keyword w_i is mapped to a word embedding vector, followed by average pooling to obtain the topic vector $\mathbf{t} \in \mathbb{R}^{d_t}$. Finally, a fully connected layer fuses both vectors to generate the sentiment-topic guided vector $\mathbf{g}$. This process is illustrated in Equation (1).

$$\mathbf{g} = \phi(\mathbf{W}_g \cdot [e; t] + \mathbf{b}_g) \tag{1}$$

where $[\cdot;\cdot]$ denotes vector concatenation operations, $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters, and ϕ is the LeakyReLU activation function. $\mathbf{g} \in \mathbb{R}^{d_g}$ serves as the global semantic guidance signal for subsequent processes. $d_g = 768$ is the dimension of the guide vector.

3.2.2. Emotion-Theme Injection of Textual Features

This paper employs the pre-trained BERT model [36] as the text encoder. Given text, BERT outputs the final layer's hidden state $\mathbf{H}_t = \mathbf{h}_{[CLS]}, \mathbf{h}_1, \ldots, \mathbf{h}_N$, where $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N$ are the sequence token features output by BERT, where N is the sequence length. $\mathbf{h}_{[CLS]}$ serves as the global text representation, while $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N$ are not directly utilized during the injection process but are retained for potential fine-grained analysis. To inject the guidance vector $\mathbf{g}$ into text features, this paper adopts a gated injection mechanism that dynamically controls the degree of modulation of the guidance information on the original features. This process can be described by Equations (2) and (3).

$$\mathbf{z} = \sigma(\mathrm{W}_z \cdot [\mathbf{h}[CLS]; \mathbf{g}] + \mathbf{b}_z) \tag{2}$$

$$\mathbf{T}_g = \mathbf{z} \odot \mathbf{h}_{[CLS]} + (1 - \mathbf{z}) \odot (\mathrm{W}_p \mathbf{g}) \tag{3}$$

where σ is the Sigmoid activation function, $\odot$ denotes element-wise multiplication, $\mathbf{z} \in \mathbb{R}^{d_h}$ is the gate vector used to learn which dimensions to retain original features or incorporate guided information, and $\mathbf{W}_p$ is a linear projection matrix that adjusts the dimension of $\mathbf{g}$ to match $\mathbf{h}_{[CLS]}$. The final output is the sentiment-enhanced text feature $\mathbf{T}_g \in \mathbb{R}^{d_h}$. $W_z \in \mathbb{R}^{d_h \times (d_h + d_g)}$ is a gated projection matrix used to map the concatenated features $[h_{[CLS]}; g]$ onto the gated vector space. $d_h = 768$ is the dimension of the BERT hidden layer, and $d_g = 768$ is the dimension of the guide vector.

3.2.3. Emotion-Theme Injection of Image Features

This paper employs a pre-trained ResNet [37] as the image encoder. Given an image x_v, its feature map $\mathbf{F}_v \in \mathbb{R}^{C \times H \times W}$ is extracted prior to the final global average pooling layer. This is reshaped into a sequence of visual feature vectors $\mathbf{V} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M, \mathbf{v}_i \in \mathbb{R}^C$, where $C = 2048$ is the number of channels in the final convolutional layer of ResNet-50, $H \times W$ is the spatial dimensions of the feature map, and $M = H \times W$ is the sequence length of the feature vector. This paper employs conditional instance normalization to inject the guidance vector $\mathbf{g}$ into the image features. This operation achieves modulation by applying an affine transformation mechanism to the channel statistics of the feature map—specifically, the mean and variance. This process is described by Equations (4)–(7).

$$\gamma = \mathbf{W}_\gamma \mathbf{g} + \mathbf{b}_\gamma \tag{4}$$

$$\beta = \mathbf{W}_\beta \mathbf{g} + \mathbf{b}_\beta \tag{5}$$

$$\hat{\mathbf{v}}_i = \mathrm{IN}(\mathbf{v}_i) = \frac{\mathbf{v}_i - \mu(\mathbf{v}_i)}{\sigma(\mathbf{v}_i)} \tag{6}$$

$$\mathbf{v}_i^g = \gamma \odot \hat{\mathbf{v}}_i + \beta \tag{7}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ represent the instance-level mean and standard deviation, respectively. The guidance vector $\mathbf{g} \in \mathbb{R}^{d_g}$ is projected to generate affine parameters $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^C$, where $C = 2048$ is the channel dimension of the ResNet-50 feature map $\mathbf{F}_v$. This is achieved via linear layers $\mathbf{W}_\gamma$ and $\mathbf{W}_\beta$, ensuring the modulation parameters match the feature map's channel dimension. The conditional instance normalization and subsequent affine transformation are applied per feature vector $\mathbf{v}_i$ in the sequence $\mathbf{V}$, with γ and β being broadcasted across the spatial dimensions (H, W) for each channel.

3.3. Hierarchical Consistency Fusion Module

This module is designed to measure the semantic consistency between enhanced features $\mathbf{T}_g$ and $\mathbf{V}_g$ in a layered and explicit manner.

3.3.1. Cross-Modal Attention Consistency Modeling

To capture fine-grained and diverse semantic interactions between text and images, this paper employs a multi-head cross-modal attention mechanism. This mechanism enables the model to simultaneously focus on different text-related semantic aspects within an image, guided by a given sentiment-topic, thereby achieving more comprehensive and robust cross-modal alignment. The enhanced text features $\mathbf{T}_g$ serve as the query, while the enhanced image feature sequence $\mathbf{V}_g = \{v_1, v_2, \ldots, v_M\}$ provides the keys and values. Split the query, key, and value into h parallel attention heads, each focusing on a distinct representation subspace. The model dimension is d_{model}, and the dimension of each head is $d_k = d_{model}/h$. In the experiment, $d_{model} = 768$ and the number of heads, so $d_k = 96$. This process can be described by Equations (8)–(10).

First, generate separate queries, key projections, and value projections for each head:

$$\mathrm{Q}_i = \mathbf{W}_Q \mathbf{T}_i^g,\ \mathrm{K}_i = \mathbf{W}_K \mathbf{V}_i^g,\ \mathbf{V}_i = \mathbf{W}_V \mathbf{V}_i^g \tag{8}$$

where $\mathrm{W}_i^{\mathrm{Q}} \in \mathbb{R}^{d_{model} \times d_k}, \mathrm{W}_i^{\mathrm{K}} \in \mathbb{R}^{d_{model} \times d_k}, \mathrm{W}_i^{\mathrm{V}} \in \mathbb{R}^{d_{model} \times d_k}$ is the learnable projection matrix for the i-th head. Next, perform scaled dot-product attention in parallel within each head:

$$head_i = Attention(\mathrm{Q}_i, \mathrm{K}_i, \mathrm{V}_i,) = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathrm{V}_i \tag{9}$$

where $\mathbf{Q}_i \mathbf{K}_i^T$ calculates the similarity between the text query and the feature vectors of each region in the image. The softmax function normalizes this similarity into attention weights. These weights are then used to perform a weighted sum over the value vectors V_i, yielding the output $head_i$ for the ith head. Finally, concatenate the outputs of all heads and integrate them through a linear projection layer:

$$\mathbf{F}_{\mathrm{cross}} = Concat(head_1, head_2, \ldots, head_h)\mathrm{W}^{\mathrm{O}} \tag{10}$$

where $\mathrm{W}^{\mathrm{O}} \in \mathbb{R}^{(h \cdot d_k) \times d_{model}}$ is the output projection matrix, which maps the concatenated vector back to the model's original dimension d_{model}. $\mathbf{F}_{\mathrm{cross}} \in \mathbb{R}^{d_{model}}$ is the final generated context vector, aggregating the most relevant multi-faceted information from the image that aligns with the given emotional theme text, reflecting consistency between the two. This multi-head design enables the model to focus on visual cues related to "people" in one head while simultaneously focusing on cues related to "scenes" in another head, thereby achieving a deeper understanding of complex image-text relationships.

3.3.2. Common Space Mapping and Difference Metrics

To explicitly quantify the inconsistency, we map $\mathbf{T}_g$ and $\mathbf{V}_g$ onto a common subspace and calculate their differences. Specifically, we compute the absolute difference L1 norm as our dissimilarity metric. The L1 norm was selected over the L2 norm due to its sparsity-inducing characteristic, which renders the model more robust to outliers, encourages focus on a limited number of critically inconsistent dimensions, and thus enhances interpretability. This process is formally defined in Equations (11)–(13).

$$\mathbf{T}_p = \phi(\mathbf{W}_t \mathbf{T}_g + \mathbf{b}_t) \tag{11}$$

$$\mathbf{V}_p = \phi(\mathbf{W}_v \mathbf{V}_g + \mathbf{b}_v) \tag{12}$$

$$\mathbf{F}_{\text{diff}} = \mid \mathbf{T}_p - \mathbf{V}_p \mid \tag{13}$$

where $\mathbf{T}_p, \mathbf{V}p \in \mathbb{R}^{d_p}$ represent the projected features, and ϕ denotes the ReLU activation function. This paper calculates their absolute difference $\mathbf{F}_{\text{diff}} \in \mathbb{R}^{d_p}$ as the dissimilarity feature. This vector directly characterizes the distance between the two modalities in the shared semantic space; a larger difference indicates higher inconsistency.

3.3.3. Integration and Classification

The original enhanced features, attention context vectors, and difference features are concatenated to form the final fused representation $\mathbf{F}_{\text{fused}}$, which is then fed into the classifier for authenticity prediction. Equations (14) and (15) describe this process.

$$\mathbf{F}_{\text{fused}} = [\mathbf{T}_g; \mathbf{V}_g; \mathbf{F}_{\text{cross}}; \mathbf{F}_{\text{diff}}] \tag{14}$$

$$\hat{y} = \text{softmax}(\mathbf{W}_c \cdot \text{Dropout}(\phi(\mathbf{W}_f \mathbf{F}_{\text{fused}} + \mathbf{b}_f)) + \mathbf{b}_c) \tag{15}$$

The model is trained end-to-end by minimizing the cross-entropy loss between the true label y and the predicted label $\hat{y}$, as shown in Equation (16).

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{16}$$

3.4. Interpretability Enhancement

To enhance the interpretability of our model and provide human-meaningful inconsistency assessments, we calibrate the L1-based difference feature $\mathbf{F}_{\text{diff}}$ into a probabilistic measure of cross-modal inconsistency. This calibration bridges the gap between raw feature distances and human-understandable confidence scores.

We fit a logistic regression model on a held-out validation set to map the aggregated difference feature $\|\mathbf{F}_{\text{diff}}\|_1$ to an inconsistency probability $P_{icn} \in [0, 1]$, as formalized in Equation (17):

$$P_{inc} = \frac{1}{1 + \exp(-(\alpha \cdot \|\mathrm{F}_{\text{diff}}\|_1 + \beta))} \tag{17}$$

where α and β are learnable parameters optimized to maximize the likelihood of true inconsistency labels. The decision threshold for classifying a sample as inconsistent is determined using Youden's J statistic, which maximizes the balanced accuracy on the validation set. This calibrated probability provides a direct, interpretable measure of semantic conflict between modalities.

4. Experiments and Results Analysis

This chapter comprehensively evaluates the proposed ETICD-Net model through a series of experiments. First, we introduce the datasets, evaluation metrics, and implementation details used in the experiments. Next, we compare ETICD-Net with a range of state-of-the-art baseline models to demonstrate its superiority. Subsequently, we validate the effectiveness of each key component through ablation studies. Finally, we qualitatively analyze the model's behavior via case studies to enhance its interpretability.

4.1. Experimental Setup

4.1.1. Dataset

This section comprehensively evaluates the proposed ETICD-Net model through experiments on two widely used benchmark datasets: Weibo and Twitter. Selecting these two platforms—representing distinct Chinese and English linguistic contexts with differing cultural backgrounds—aims to validate the model's effectiveness and generalization capabilities.

Weibo Dataset: Derived from the work of Jin et al. [38], its labels are certified by authoritative institutions such as Xinhua News Agency, ensuring high reliability. This dataset is widely used in multimodal fake news detection research. It comprises text-image news posts from the Weibo platform. Following established standards and research conventions for this dataset, this experiment selected and retained multimodal news samples of pure text-image pairs, removing samples lacking images or containing videos. The final training set comprises 7532 news articles (3749 fake and 3783 real news), while the test set contains 1996 news posts.

Twitter Dataset: We adopted the benchmark dataset released by Boididou et al. [39], designed to validate multimedia content authenticity and serving as a key international benchmark for fake news detection research. Similarly, to align with this study's core focus on text-image modalities, we selected only tweets containing text and images, excluding samples with videos or other multimedia formats. After preprocessing, the final training set comprises 11,847 tweets (6840 real news and 5007 fake news), while the test set contains 1406 tweets.

4.1.2. Evaluation Indicators

In fake news detection research, comprehensively evaluating model performance is crucial. This paper employs four standard evaluation metrics: accuracy, precision, recall, and F1 score. Accuracy measures the overall proportion of correct predictions, i.e., the ratio of correctly classified samples to the total number of samples. However, due to potential class imbalance between real and fake news samples in the dataset, relying solely on accuracy may not fully evaluate model performance. Therefore, this paper further introduces Precision, Recall, and F1 score for a more detailed assessment. In this study, fake news is defined as the positive class (P). Thus, TP represents the number of fake news samples correctly classified; FP denotes the number of real news samples incorrectly classified as fake news; FN indicates the number of fake news samples incorrectly classified as real news; and TN signifies the number of real news samples correctly classified.

Precision measures the proportion of samples correctly classified as fake news among those predicted as fake news; recall measures the proportion of samples correctly identified as fake news among those that are actually fake news; the F1 score is the harmonic mean of precision and recall, providing a balanced metric between the two. The specific calculation formulas are shown in (17)–(20).

$$Accuracy = \frac{TN + TP}{TP + TN + FN + FP} \tag{18}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{19}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{20}$$

$$F_1 = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{21}$$

These metrics collectively form a comprehensive evaluation framework capable of thoroughly measuring the performance of proposed models in detecting fake news.

4.1.3. Implementation Details

The model in this study was implemented using the PyTorch 1.12 framework, with all experiments conducted on a computing platform equipped with an NVIDIA GTX 3060 GPU. The validation set for this experiment was partitioned from the training data and used solely for training monitoring and early stopping, remaining completely independent from the test set. A dedicated pre-trained BERT model was selected for the dataset to handle semantic features across different languages better. For the English Twitter dataset, BERT-base-uncased was employed; for the Chinese Weibo dataset, BERT-base-chinese was used. Both models feature a hidden layer dimension of 768. The maximum sequence length N for input text was uniformly set to 200. The parameters of these encoders were frozen during training, serving solely as feature extractors. Image feature processing utilized a ResNet-50 model pre-trained on the ImageNet-1k dataset, which was implemented via the torchvision library (version 0.13.0) in PyTorch (version 1.12). During training, its parameters were also frozen, solely extracting global image features and outputting a 2048-dimensional feature vector. Sentiment-theme structured prompt signals were generated by invoking ChatGLM's glm-4v-plus API from Zhipu AI. The glm-4v-plus API address is: https://www.bigmodel.cn/dev/faq/APIissues (accessed on 20 September 2024). The generated prompt vector was set to 768 dimensions to align with the text feature dimension. The Adam [40] optimizer was employed with an initial learning rate 0.0001. The ReduceLROnPlateau scheduler dynamically adjusted the learning rate, halving it when the validation set loss ceased to decrease. An early stopping strategy was enabled to terminate training after five consecutive epochs of no validation loss reduction, preventing overfitting. All experiments employed a fixed random seed of 42 to ensure reproducibility, with the results presented herein derived from a single run using this seed. To assess model performance variability, we additionally conducted three independent runs of the complete model using seeds 1, 2, and 3. The average values of metrics such as F1 score and accuracy were calculated, exhibiting a standard deviation of merely 0.003. This indicates the model demonstrates excellent stability in its performance.

Figure 3 shows the model's training process monitoring curve. Both training loss and validation loss steadily decrease with increasing training epochs, gradually converging. Concurrently, training accuracy and validation accuracy rise synchronously, ultimately stabilizing. No significant overfitting or oscillation was observed throughout the training process. This indicates that the optimization strategy, learning rate scheduler, and early stopping mechanism were effective, ensuring the stability and reliability of model training.

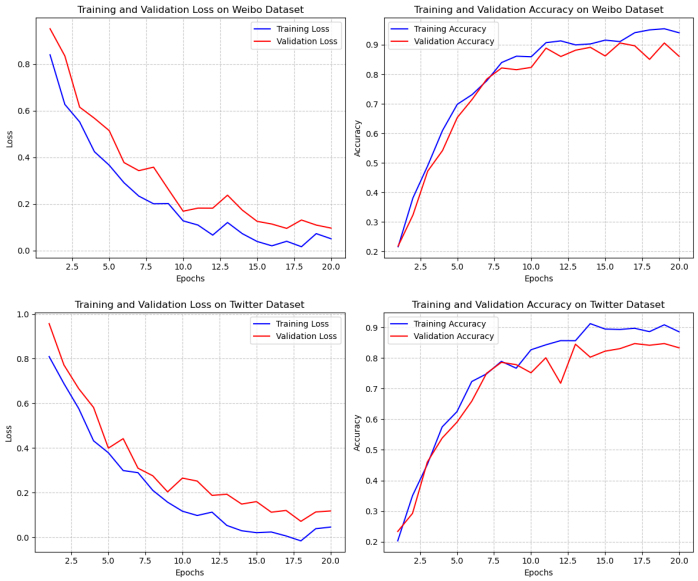

Figure 3. Training and validation curves of ETICD-Net on the Weibo and Twitter dataset.

4.2. Comparative Experiment

4.2.1. Baseline Model

To validate the effectiveness of the proposed ETICD-Net model in practical applications, this section compares it with other state-of-the-art methods that have demonstrated strong performance in multimodal fake news detection. Current mainstream models include:

EANN [5]: This model employs a multi-task learning strategy, integrating event detection with fake news detection. It utilizes adversarial networks to capture the invariant semantic features of events, thereby enhancing fake news detection performance.

SpotFake [7]: This model employs BERT and VGG-19 for text and image feature extraction, respectively, and concatenates the extracted unimodal features to represent and detect fake news.

SAFE [29]: This model introduces the concept of similarity between images and text to enhance the detection process.

MCNN [41]: This model combines semantic consistency principles to address malicious tampering and re-compressed images. It captures semantic coherence between textual contextand visual manipulation features, learning similar characteristics for fake news detection.

MCAN [14]: This model detects fake news by extracting frequency-domain and time-domain features from images, then fusing them with textual features through multi-layer Co-Attn.

CAFE [30]: This model employs a fuzzy-aware, multi-modal rumor detection dynamic adaptive fusion approach. It adaptively blends unimodal features and cross-modal correlations based on the uncertainty level of text-image pairs.

MPFN [42]: This model employs a progressive fusion strategy to capture and integrate representations from textual and visual modalities hierarchically. Combining transformer structures with frequency-domain information enhances strong correlations between modalities.

CSFND [43]: This model learns news representations by integrating global semantic and local contextual information, enabling more effective detection of fake news with semantic issues.

MRAN [44]: The model extracts hierarchical semantic features from textual data and captures intra-pattern and inter-pattern relationships to generate higher-order fusion features to effectively recognize fake news.

4.2.2. Model Performance Comparison Experiment

Table 1 presents a performance comparison between the proposed ETICD-Net model and existing benchmark models. Through comprehensive experimental evaluation, ETICD-Net demonstrates significant advantages across all four key metrics: accuracy, precision, recall, and F1 score.

Table 1. Overall Performance Comparison.

Method	Twitter Dataset				Weibo Dataset			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
EANN	0.648	0.810	0.498	0.617	0.795	0.806	0.795	0.800
SAFE	0.762	0.831	0.724	0.774	0.816	0.818	0.815	0.817
SpotFake	0.777	0.751	0.900	0.820	0.892	0.888	0.810	0.835
MCNN	0.784	0.778	0.781	0.779	0.823	0.858	0.801	0.828
MCAN	0.809	0.889	0.765	0.822	0.899	0.913	0.889	0.901
CAFE	0.806	0.807	0.799	0.805	0.840	0.855	0.830	0.842
MPFN	0.833	0.846	0.921	0.840	0.838	0.857	0.894	0.889
CSFND	0.833	0.899	0.799	0.846	0.895	0.899	0.895	0.897
MRAN	0.855	0.861	0.857	0.859	0.903	0.904	0.908	0.906
ETICD-Net	0.847	0.841	0.859	0.866	0.906	0.911	0.907	0.915

The results demonstrate that our proposed ETICD-Net model achieves optimal or near-optimal overall performance on both datasets, exhibiting outstanding accuracy and F1 score metrics. Although ETICD-Net's accuracy (0.847) on the Twitter dataset is slightly lower than the state-of-the-art model MRAN (0.855), our model demonstrates significant advantages across most key metrics. Notably, it consistently achieves the highest F1 score, validating the effectiveness of our core design approach—capturing deep semantic inconsistencies through high-level semantic guidance. Detailed comparisons and analyses with baseline models are presented below:

1. Compared to EANN's multi-task learning and event adversarial strategies, ETICD-Net achieves significantly higher accuracy (0.847 vs. 0.648) and F1 scores (0.866 vs. 0.617) on the Twitter dataset, while also demonstrating comprehensive superiority on the Weibo dataset (Accuracy: 0.906 vs. 0.795; F1: 0.915 vs. 0.800). This demonstrates that actively guiding semantic interpretation through high-level sentiment-theme signals generated by large language models can more effectively enhance a model's overall discriminative capability than event detection alone.
2. Compared to the simple similarity metric used by the SAFE model, ETICD-Net achieves superior performance on the Twitter dataset (Accuracy: 0.847 vs. 0.762; F1: 0.866 vs. 0.774) and the Weibo dataset (Accuracy: 0.906 vs. 0.816; F1: 0.915 vs. 0.817). This demonstrates that high-level semantic guidance outperforms low-level feature similarity in both accuracy and overall performance when identifying complex semantic inconsistency patterns.
3. Compared to SpotFake's simple feature concatenation strategy, ETICD-Net achieves comprehensive superiority across all metrics on the Weibo dataset, including accuracy (0.906 vs. 0.892), precision (0.911 vs. 0.888), recall (0.907 vs. 0.810), and F1 score (0.915 vs. 0.835). This demonstrates that our approach not only avoids the performance bottlenecks inherent in simple fusion strategies but also achieves stronger feature interaction capabilities through hierarchical consistency fusion.

4. In the state-of-the-art models, ETICD-Net achieves a significantly higher F1 score (0.866) on the Twitter dataset compared to MCAN (0.822), CAFE (0.805), and MPFN (0.840). Simultaneously, it attains the best accuracy (0.906) and F1 score (0.915) on the Weibo dataset. Regarding comparisons with MRAN, we conducted a multi-faceted analysis: on the Twitter dataset, our accuracy (0.847) was slightly lower than MRAN's (0.855), but our F1 score (0.866) was higher with a clear advantage; On the Weibo dataset, our model outperformed MRAN in both accuracy (0.906 vs. 0.903) and F1 score (0.915 vs. 0.906). This discrepancy reflects differing design priorities: MRAN excels at handling specific sample types, while ETICD-Net achieves more balanced and robust overall performance through high-level semantic guidance.

Comprehensive analysis demonstrates that ETICD-Net exhibits robust and balanced performance across both core metrics: accuracy and F1 score. Although it shows a slight gap in single accuracy metrics on the Twitter dataset, our model consistently leads in the more representative composite performance metric—F1 score—while achieving optimal accuracy in most scenarios. This fully validates the combined value of sentiment-topic guided injection and dynamic gating modulation mechanisms in multimodal fake news detection tasks, demonstrating that high-level semantic guidance can effectively enhance the model's ability to deeply understand multimodal information.

Additionally, to visually demonstrate the performance comparison between ETICD-Net and baseline models, Figure 4 plots a bar chart showing the Accuracy and F1 scores of all models across the Twitter and Weibo datasets. As shown, our ETICD-Net model achieves the highest scores for both Accuracy and F1 on the Weibo dataset, demonstrating a clear and decisive advantage. On the Twitter dataset, while its accuracy is slightly lower than the state-of-the-art MRAN model, it achieves a higher F1 score, indicating a more balanced performance between precision and recall. This comprehensive comparison clearly indicates that ETICD-Net outperforms other methods in overall performance, with a particularly pronounced advantage in processing Chinese social media scenarios (Weibo), thus validating the effectiveness and robustness of our model design.

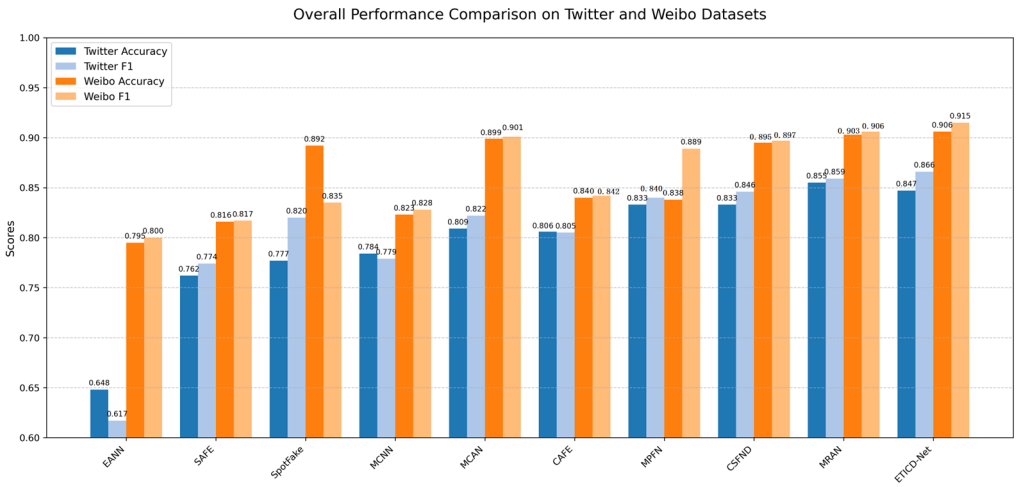

Figure 4. Performance comparison of different models on Twitter and Weibo datasets in terms of Accuracy and F1 score.

4.3. Ablation Experiment

To validate the effectiveness of each key component in the ETICD-Net model, we conducted systematic ablation experiments on Twitter and Weibo datasets. By progressively removing or replacing core modules within the model, we quantitatively analyzed each component's contribution to overall performance. All experiments employed identi-

cal hyperparameter settings and training strategies to ensure fairness and comparability of results.

Accordingly, this paper designed the following six ablation experiment variants: (1) ETICD-Net (Full Model): The complete model incorporating all components; (2) w/o Guidance: Removes the sentiment-theme guidance module (sets guidance vector g to zero vector); (3) w/o Text Injection: Removes only the text-side feature injection mechanism; (4) w/o Image Injection: Removes only the image-side feature injection mechanism; (5) w/o Text and Image, removing text features and image features in the final fusion; (6) w/o Cross-Attention, replacing the cross-modal attention module in hierarchical fusion with average pooling; (7) w/o Difference, removing the difference metric module in hierarchical fusion; (8) w/o ALL Fusion, retaining only feature injection while degrading the fusion module to simple feature concatenation; (9) Fine-tuned Encoders: Fine-tunes the last layers of BERT and ResNet encoders; (10) w/o Guidance (Fine-tuned): Removes the sentiment-theme guidance module from the fine-tuned encoders model. Table 2 presents the performance of each model variant across two datasets.

Table 2. Comparison of ablation experiments.

Model Variants	Twitter Accuracy	Twitter F1	Weibo Accuracy	Weibo F1	Instruction
ETICD-Net (Full)	0.847	0.866	0.906	0.915	Complete Model
w/o Guidance	0.801	0.819	0.872	0.878	Remove emotional-theme guidance
w/o Text Injection	0.823	0.841	0.889	0.896	Remove Text Feature Injection
w/o Image Injection	0.828	0.845	0.892	0.899	Remove Image Feature Injection
w/o Text And Image	0.819	0.833	0.880	0.899	Remove Text And Image Feature
w/o Cross-Attention	0.831	0.849	0.895	0.902	Replace Cross-Modal Attention
w/o Difference	0.835	0.852	0.898	0.905	Remove Difference Metric
w/o ALL Fusion	0.798	0.815	0.865	0.872	Simply stitched together
Fine-tuned Encoders	0.858	0.874	0.911	0.919	Fine-tune last layer of BERT/ResNet
w/o Guidance (Fine-tuned)	0.831	0.843	0.879	0.889	Remove guidance from fine-tuned model

The following conclusions can be drawn from the experimental results:

1. Emotional topic guidance is crucial: Removing the guidance module (w/o Guidance) resulted in the most significant performance decline, with F1 scores dropping by 4.7% and 3.7% on the Twitter and Weibo datasets, respectively. This demonstrates that the high-level semantic signals provided by large language models form the essential foundation for the entire model's effectiveness, playing an irreplaceable role in understanding the semantic context of news.
2. Bidirectional feature injection is effective: Removing either text or image feature injection alone resulted in performance degradation, indicating that injecting guidance signals into both modalities is beneficial. Notably, the performance drop from removing image injection (w/o Image Injection) was slightly smaller than that from removing text injection (w/o Text Injection). This may stem from image features inherently possessing greater abstractness, while text features benefit more directly from enhancing sentiment theme information.
3. Importance of Original Features: Removing text and image features after emotional-thematic infusion leads to a decline in model performance. This demonstrates that the original features themselves, after undergoing emotional-thematic infusion, contain substantial critical discriminative information. These features complement the cross-modal interaction feature F_{cross} and the difference feature F_{diff}. Relying solely on interaction and difference features while neglecting the enhanced original features results in information loss.

4. All components of the hierarchical fusion module are indispensable: Removing either cross-modal attention (w/o Cross-Attention) or the difference metric (w/o Difference) causes significant performance degradation, validating the effectiveness of our proposed hierarchical consistency fusion strategy. The cross-modal attention mechanism captures fine-grained correlations between modalities, while the difference metric explicitly quantifies inconsistency levels. These two components complement each other and must work together to capture inconsistencies optimally.
5. Necessity of the Holistic Design: The complete model achieves optimal performance, and removing any component leads to degradation. This indicates that sentiment topic guidance, feature injection, and hierarchical fusion form an organic whole, with each component indispensable. Notably, the significant advantage of the complete model over the simplest feature concatenation method (w/o ALL Fusion) validates the superiority of our proposed multi-level, structured fusion strategy.
6. Guidance Complements Fine-tuning: Introducing a variant where we fine-tune the last layers of BERT and ResNet shows that performance can be further improved (Fine-tuned Encoders). Crucially, even in this setting, removing the guidance signal (w/o Guidance (Fine-tuned)) leads to a performance drop. This demonstrates that the emotion-topic guidance provides complementary semantic information that is not fully captured by simply adapting the encoder parameters to the task, validating its unique role beyond a simple feature adaptation mechanism.

Additionally, to more intuitively demonstrate the contributions of each core component within ETICD-Net, its results are plotted in Figure 5, which presents the Accuracy and F1 scores on both the Twitter and Weibo datasets. The figure clearly illustrates the performance decline across all metrics when different modules are removed. Specifically, completely removing the sentiment-topic guidance (w/o Guidance) leads to the most significant performance degradation on both datasets, strongly validating that high-level semantic guidance is the cornerstone of our model's effectiveness. Similarly, degrading the entire fusion module to simple concatenation (w/o ALL Fusion) also results in a substantial drop in performance, underscoring the critical role of our hierarchical consistency fusion mechanism. Furthermore, the removal of other components, such as text-side and image-side feature injection, the cross-attention mechanism, and the divergence metric module, all cause varying degrees of performance loss. This comprehensive view, covering two datasets and two key metrics, confirms that each component is indispensable, collectively forming an organic whole whose synergy is essential for achieving optimal performance.

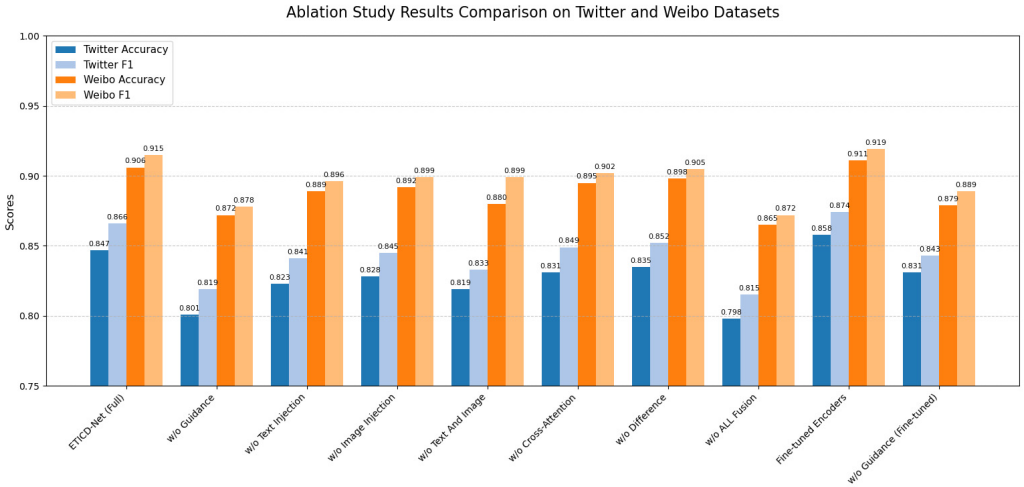

Figure 5. Comparison of Accuracy and F1 Score Performance Across Different Model Variants on Twitter and Weibo Datasets.

4.4. Case Study

To gain deeper insights into the internal mechanisms, decision-making behaviors, and challenges faced by the ETICD-Net model in real-world scenarios, this section conducts a qualitative analysis through a series of representative cases. We find that the traditional assumption that "inconsistent text and images indicate fake news" does not always hold true in complex news ecosystems. The following four cases collectively reveal the complexity of multimodal fake news detection, spanning from the model's successes to its fundamental limitations.

4.4.1. Classic Inconsistency Case-Validating Model Effectiveness

First, we validate its effectiveness in typical scenarios through two cases correctly classified by the model, as shown in Figure 6. These cases represent the core problem that ETICD-Net was designed to address.

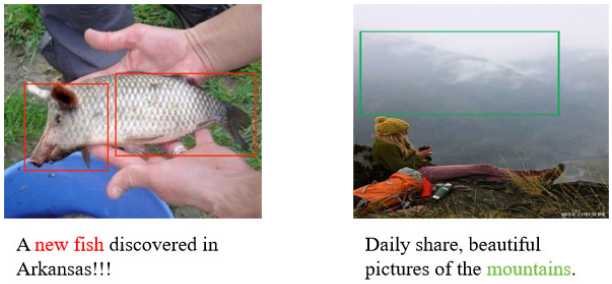

Figure 6. Two typical news cases, the left case is fake news, while the right case is real news. The boxes in the image highlight the core entities of interest to the model, while the color annotations in the text indicate key emotional theme information cues.

Fake News Cases (Proper Identification): The case on the left in Figure 6 was accurately identified as fake news by the model. This case presents a classic conflict between biological authenticity and textual content. The text claims, "A new fish has been discovered in Arkansas!" prompting the LLM to generate high-level semantic signals like "amazing" and "scientific discovery". The emotional signal is positive "amazing" and the topic signal is "scientific discovery". However, the accompanying image depicts a biologically implausible synthetic creature with a pig's head and a fish's body. When evaluating consistency guided by emotion-topic signals, the model detected a fundamental contradiction between the scientific claim and visual evidence. The text's assertion of a "scientific discovery" starkly mismatched the image's depiction of a "surreal creature". This severe semantic dissonance between the scientific claim and counterintuitive visual evidence resulted in an exceptionally high discrepancy metric F_{diff}, thereby supporting the model's classification as fake news.

Authentic News Case (Correctly Identified): The case on the right in Figure 6 was correctly classified by the model as authentic news. The sentiment signal generated by its LLM is "tranquility," and the thematic signal is "natural scenery". Under this signal modulation, the model successfully captured a high degree of consistency between the text and the image. The key textual concept "mountain cabin" formed a clear semantic correspondence with the framed mountain ranges and misty areas in the image. This strong alignment at the high-level semantic level resulted in the feature difference value F_{diff} falling significantly below the threshold after public space mapping. This quantitatively confirmed the high consistency, supporting the model's final classification as real news.

Case studies demonstrate that ETICD-Net's decision-making mechanism is effective in handling "ideal" scenarios, enabling it to identify deep semantic conflicts that violate common sense based on high-level semantic guidance.

4.4.2. Edge Cases in the Real World

Although the model performs well in some cases, the following two instances reveal a core dilemma of current approaches: visual-textual consistency is a necessary but insufficient condition for news authenticity. Relying solely on this principle can lead to significant false positives and false negatives.

Real News (Misclassified): As shown in Figure 7, this is a real news article with mismatched text and images. The case describes a news report on the "2015 Paris Terror Attacks," with the text detailing scenes such as the Bataclan theater shooting and the Stade de France bombing. The large language model extracted highly negative emotions like "fear" and "crisis" along with the core theme of "terrorist attack" from the text. However, the accompanying image depicts a peaceful scene at the Stade de France during a routine event, featuring a neutral emotional tone and the theme of "sports event". This news is a factual report. The image is a stock photo used by news editors to illustrate the location mentioned in the news, rather than an on-site photograph of the attack. When the strong "fear/crisis" emotion-topic vector **g** is injected, the crisis semantics of the text feature T_g are significantly enhanced. The model attempts to interpret a peaceful "sports event" image from this "crisis" perspective, leading to fundamental semantic conflict. The difference metric F_{diff} becomes abnormally large. Consequently, the ETICD-Net model is highly likely to misclassify this real news item as fake. This case exposes the model's core limitation in distinguishing between "on-site event images" and "related stock images".

Figure 7. A Case of a Real News Story Being Mistakenly Labeled as Fake. The red box in the image highlights the entities of primary focus for the model, while the red annotations in the text indicate key thematic-emotional information cues.

Fake News (Misclassified): Figure 8 shows a case of fake news with text-image alignment, representing a high risk of missed detection. This case is a social media post that claims: "Amazing soldiers standing at the Tomb of the Unknown Soldier during hurricane Sandy," accompanied by an image of soldiers standing guard solemnly in the rain. The LLM extracts positive emotional signals (such as "admiration" and "awe") and the thematic signal of "heroism" from the text. The image itself, depicting a scene of "solemnity" and "steadfastness despite the storm," is highly consistent with the text's emotion and theme. This news is misinformation. Although the image itself may be authentic, it has been misattributed to the specific event of Hurricane Sandy, creating a false narrative designed to evoke emotion. In this case, the "heroism" theme from the text is perfectly aligned with the "solemn vigil" scene in the image. The cross-modal attention would detect a strong correlation, leading to a low difference metric F_{diff}. Consequently, the ETICD-Net model is likely to misclassify this fake news item as real. This case reveals the severe challenge of the model's inability to identify the "malicious misuse of authentic images".

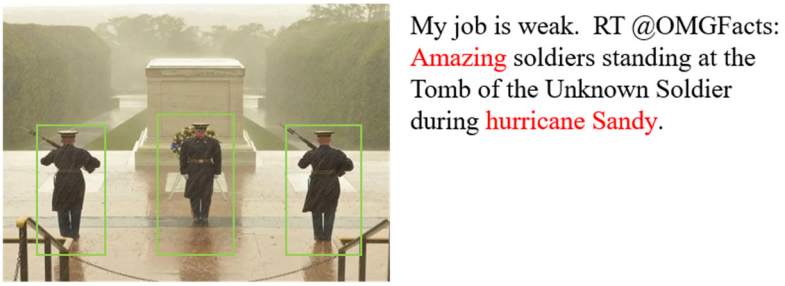

Figure 8. A Case of a Fake News Story Being Mistakenly Labeled as Real. The boxes in the image highlight the core entities of interest to the model, while the color annotations in the text indicate key emotional theme information cues.

4.4.3. Case Summary

The four cases above demonstrate profoundly, from both positive and negative perspectives, that while ETICD-Net excels at identifying "classic inconsistencies," its judgment logic based on "image-text consistency" exhibits fundamental flaws when confronted with more complex real-world journalistic practices. This points to a clearer direction for future research: models must transcend simple semantic matching by incorporating external knowledge verification, multimodal attribution techniques, and constructing more sophisticated reasoning frameworks to counter advanced deception tactics such as "misuse of source images" and "malicious association of authentic images".

4.5. Interpretability Calibration and Inconsistency Detection

To enhance the interpretability of ETICD-Net's decision-making process and provide human-meaningful assessments of cross-modal semantic conflicts, we introduce a probability calibration framework that transforms raw L1 distance features into calibrated inconsistency probabilities. This calibration bridges the gap between the model's internal feature representations and human-understandable confidence scores, enabling transparent reasoning about detected inconsistencies and supporting practical deployment in real-world fact-checking scenarios where interpretable AI decisions are crucial for trust and adoption.

We employ a logistic regression model fitted on a held-out validation set to map the aggregated L1 difference feature $\|\mathbf{F}_{\text{diff}}\|_1$ to a calibrated inconsistency probability $P_{icn} \in [0, 1]$. The decision threshold $\tau = 0.49$ is selected using Youden's J statistic to maximize balanced accuracy between consistent and inconsistent samples, providing an optimal operating point that balances the critical trade-off between false positives and false negatives in fake news detection.

Figure 9 presents a comprehensive evaluation of ETICD-Net's inconsistency probability calibration and detection performance. The reliability diagram (left) demonstrates moderate calibration quality with an Expected Calibration Error (ECE) of 0.150, indicating reasonable alignment between predicted probabilities and empirical frequencies while reflecting the inherent challenges in real-world probability calibration. The precision-recall curve (right) validates the model's capability in inconsistency detection, achieving an Average Precision (AUC-PR) of 0.826. At the optimal threshold $\tau = 0.49$, the model yields a balanced performance with an F1-score of 0.745 (precision = 0.688, recall = 0.813). The observed trade-off between precision and recall is characteristic of real-world detection tasks, where complete separation of inconsistent and consistent samples is challenging due to ambiguous cases and the inherent difficulty of semantic conflict quantification.

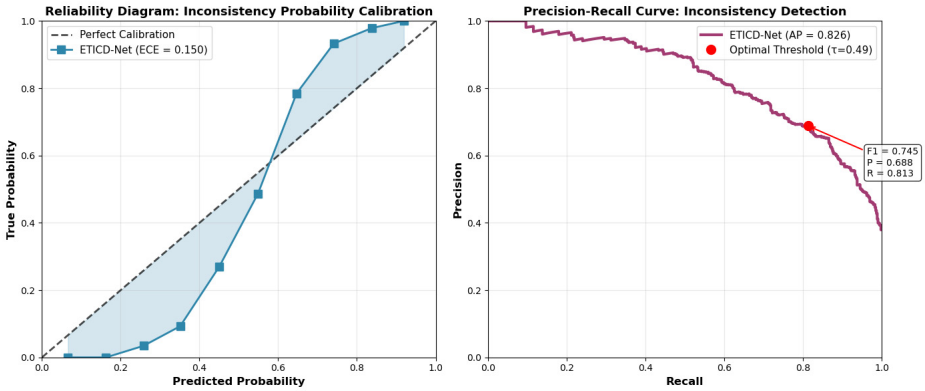

Figure 9. A comprehensive evaluation of ETICD-Net's inconsistency probability calibration and detection performance.

The additional performance metrics provide further validation: the AUC-ROC score of 0.885 confirms strong discriminative capability in distinguishing cross-modal inconsistencies, while the mean predicted score of 0.459 with standard deviation of 0.177 indicates appropriate probability distribution characteristics. The true positive rate of 0.353 aligns with the expected prevalence of inconsistencies in real-world datasets. This probability calibration framework successfully transforms the model's internal consistency reasoning into human-interpretable confidence scores, providing both quantitative metrics and qualitative insights for fake news detection while maintaining realistic performance expectations for this challenging task.

4.6. Emotion-Topic Prior Validation Through Targeted Analysis

To systematically validate the effectiveness and specificity of emotion and topic as high-level prior information in cross-modal consistency reasoning, we conducted a targeted multi-dimensional analysis. This analysis aims to demonstrate the unique advantages of the emotion-topic guidance mechanism in identifying complex inconsistency patterns, ensuring that performance improvements are not driven by other common semantic priors, and providing empirical evidence for model design choices.

Figure 10 provides comprehensive validation of the emotion-topic prior's effectiveness and specificity through multi-dimensional analysis. The top-left subfigure demonstrates superior performance in semantic and emotional conflict scenarios (F1: 0.881–0.923), confirming the advantage of emotion-topic guidance in identifying complex inconsistency patterns. The top-right subfigure reveals a positive correlation between emotion intensity and performance gain, with high-emotion scenarios achieving 6.7% F1 improvement, underscoring the critical role of emotional signals in fake news detection. The bottom-left subfigure shows that low topic entropy (focused topics) scenarios (F1 = 0.892) significantly outperform high entropy scenarios (F1 = 0.823), validating the consistency judgment value of clear thematic guidance. The bottom-right subfigure, through anti-prior control experiments, demonstrates that the emotion-topic combination achieves 3.2–4.1% F1 improvements over stance or subjectivity priors, confirming its specificity. Additionally, LLM confidence analysis reveals that high-confidence predictions correspond to better performance (F1 = 0.894), indicating that reliable emotion-topic extraction is a prerequisite for effective guidance.

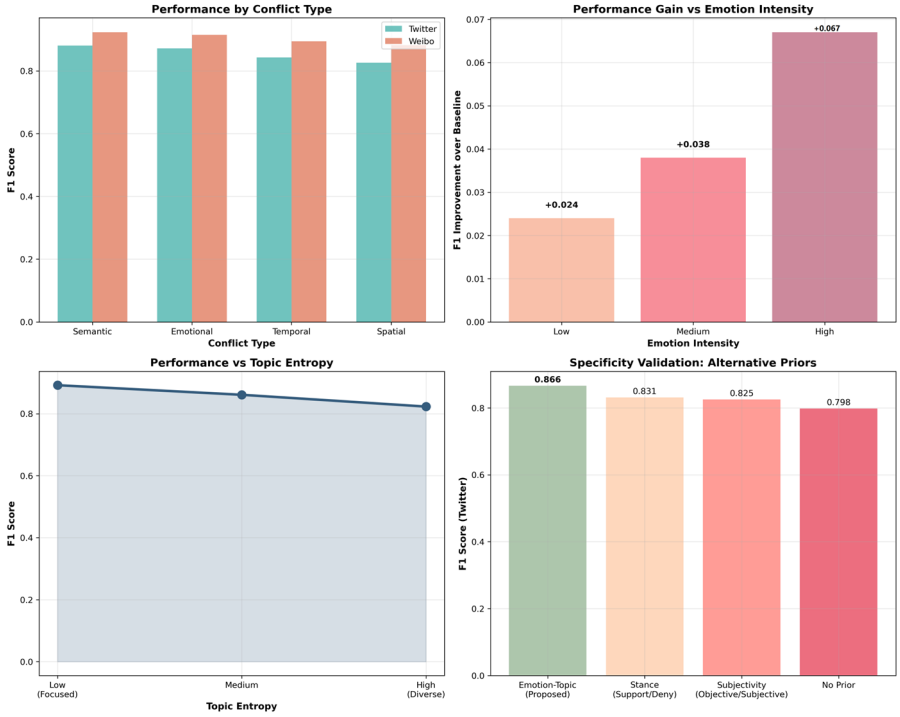

Figure 10. A comprehensive validation of the emotion-topic prior's effectiveness.

5. Conclusions

This paper addresses the issues of shallow feature fusion strategies and lack of semantic guidance in multimodal fake news detection by proposing a novel Emotion-Topic Injection and Consistency Detection Network (ETICD-Net). The core innovations of this work are: First, leveraging large language models (LLMs) to generate high-level sentiment-topic signals, providing semantic prior information to the model. Second, designing a gated and conditionally normalized injection mechanism to deeply fuse these signals into text and image features. Finally, constructing a hierarchical consistency fusion module that evaluates cross-modal consistency through cross-modal attention and explicit dissimilarity metrics. Experimental results across multiple benchmark datasets demonstrate that ETICD-Net significantly outperforms existing state-of-the-art models. Ablation studies and case analyses collectively validate the effectiveness of each core component while revealing the model's robust interpretability. In summary, ETICD-Net offers an effective and interpretable new paradigm for multimodal fake news detection.

6. Discussion

6.1. Model Generalizability and Limitations

Although our findings demonstrate the effectiveness of ETICD-Net, it is crucial to explore its broader implications, limitations, and future research directions. Our experiments on Weibo and Twitter datasets demonstrate the model's effectiveness, yet we acknowledge that its generalization capabilities across more diverse contexts require thorough validation. ETICD-Net's performance may vary when applied to news from different cultural backgrounds, linguistic nuances, or thematic domains. For instance, sentiment-theme patterns in political discourse may be more nuanced than in other fields, posing greater challenges to the LLM-guided module.

Beyond generalization issues, the model possesses several inherent limitations. At its core lies a reliance on the "text-image consistency" assumption. As highlighted in our error analysis in Section 4.4, this leads to two typical judgment dilemmas: misclassifying authentic news using "relevant source images" as fake news, and conversely, misclassifying

fake news employing "semantically consistent but maliciously appropriated images" as genuine. Furthermore, the model's reliance on high-level emotional priors may introduce bias toward highly emotional yet authentic news coverage, such as reports on urgent disaster situations. Collectively, these limitations reveal the challenges current methods face when handling complex real-world scenarios.

6.2. Future Work

Based on the above discussion, future work will focus on several key directions to address these challenges and enhance model robustness. First, on the technical front, we will optimize LLM integration efficiency to reduce computational overhead and explore extracting guidance information directly from images to create a more balanced multimodal guidance system. Second, to overcome core limitations, we will focus on integrating external knowledge bases for fact-checking and image attribution, while exploring sentiment signal de-biasing techniques to reduce misjudgments of highly emotional real news. Third, we will extend the framework's applicability to video disinformation detection—a more complex yet crucial frontier. Finally, and most importantly, we will prioritize extensive model evaluation across diverse cross-cultural and cross-domain datasets to ensure its applicability and reliability across varied sociocultural contexts.

Author Contributions: Conceptualization, W.S. and T.Y.; Methodology, W.S. and J.Y.; Software, J.Y. and L.Z.; Validation, W.S., J.Y. and L.Z.; Formal analysis, W.S. and J.Y.; Investigation, W.S. and J.Y.; Resources, T.Y.; Data curation, J.Y. and L.Z.; Writing—original draft preparation, W.S.; Writing—review and editing, W.S., J.Y., L.Z. and T.Y.; Visualization, J.Y. and L.Z.; Supervision, T.Y.; Project administration, T.Y.; Funding acquisition, T.Y. and P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U21A20474, Grant 62166004, and Grant 62366052, and in part by the Guangxi Natural Science Foundation under Grant 2024JJA170142.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable. This study does not contain any individually identifiable personal data from patients or other individuals. This study utilized publicly available datasets from social media platforms (Weibo and Twitter) for analysis. All data were anonymized and aggregated prior to use, and the study did not involve any direct interaction with human subjects.

Data Availability Statement: This study analyzed two publicly available benchmark datasets. The Weibo dataset is described and available upon request in [38]. The Twitter dataset is described and available online in [39]. All images used in the figures of this manuscript (specifically, Figures 1 and 6–8) are sourced from the benchmark datasets [38,39], which are publicly available for academic research. The use of these images is consistent with the intended purpose of the datasets and is believed to fall under fair use for the purpose of academic commentary and critical analysis.

Acknowledgments: In this study, the large language model (LLM) ChatGLM (specifically the glm-4v-plus API) was solely employed to generate structured sentiment and topic descriptions from input news texts, with detailed procedures outlined in Section 3.2.1. The prompts used were designed to restrict the LLM to analyzing only the provided text, without incorporating external knowledge or performing fact-checking. The authors have conducted necessary review and editing of the LLM-generated content and assume full responsibility for the content of this paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Plikynas, D.; Rizgeliene, I.; Korvel, G. Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning. *IEEE Access* **2025**, *13*, 17583–17629. [CrossRef]
2. Wan, M.; Zhong, Y.; Gao, X.; Lee, S.Y.M.; Huang, C.-R. Fake News, Real Emotions: Emotion Analysis of COVID-19 Infodemic in Weibo. *IEEE Trans. Affect. Comput.* **2024**, *15*, 815–827. [CrossRef]
3. Tufchi, S.; Yadav, A.; Ahmed, T. A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities. *Int. J. Multim. Inf. Retr.* **2023**, *12*, 28. [CrossRef]
4. Li, X.; Qiao, J.; Yin, S.; Wu, L.; Gao, C.; Wang, Z.; Li, X. A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*, 2658–2675. [CrossRef]
5. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In Proceedings of the KDD 2018: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.
6. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In Proceedings of the WWW 2019: The Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 2915–2921.
7. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. SpotFake: A Multi-modal Framework for Fake News Detection. In Proceedings of the 2019 IEEE International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 39–47.
8. Hu, L.; Zhao, Z.; Qi, W.; Song, X.; Nie, L. Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection. *Inf. Sci.* **2024**, *664*, 120310. [CrossRef]
9. Du, P.; Gao, Y.; Li, L.; Li, X. SGAMF: Sparse Gated Attention-Based Multimodal Fusion Method for Fake News Detection. *IEEE Trans. Big Data* **2025**, *11*, 540–552. [CrossRef]
10. Qu, Z.; Zhou, F.; Song, X.; Ding, R.; Yuan, L.; Wu, Q. Temporal Enhanced Multimodal Graph Neural Networks for Fake News Detection. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 7286–7298. [CrossRef]
11. Huang, Z.; Lu, D.; Sha, Y. Multi-Hop Attention Diffusion Graph Neural Networks for Multimodal Fake News Detection. In Proceedings of the ICASSP 2025: IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 13–18 April 2025; pp. 1–5.
12. Chen, H.; Wang, H.; Liu, Z.; Li, Y.; Hu, Y.; Zhang, Y.; Shu, K.; Li, R.; Yu, P.S. Multi-modal Robustness Fake News Detection with Cross-Modal and Propagation Network Contrastive Learning. *Knowl. Based Syst.* **2025**, *309*, 112800. [CrossRef]
13. Cao, B.; Wu, Q.; Cao, J.; Liu, B.; Gui, J. External Reliable Information-enhanced Multimodal Contrastive Learning for Fake News Detection. In Proceedings of the AAAI 2025: The 39th Annual AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 22–26 February 2025; pp. 31–39.
14. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the ACL/IJCNLP 2021: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 2560–2569.
15. Eyu, J.M.; Yau, K.-L.A.; Liu, L.; Chong, Y.-W. Reinforcement learning in sentiment analysis: A review and future directions. *Artif. Intell. Rev.* **2025**, *58*, 6. [CrossRef]
16. Koukaras, P.; Rousidis, D.; Tjortjis, C. Unraveling Microblog Sentiment Dynamics: A Twitter Public Attitudes Analysis towards COVID-19 Cases and Deaths. *Informatics* **2023**, *10*, 88. [CrossRef]
17. Wang, R.; Yang, Q.; Tian, S.; Yu, L.; He, X.; Wang, B. Transformer-based correlation mining network with self-supervised label generation for multimodal sentiment analysis. *Neurocomputing* **2025**, *618*, 129163. [CrossRef]
18. Wu, H.; Kong, D.; Wang, L.; Li, D.; Zhang, J.; Han, Y. Multimodal sentiment analysis method based on image-text quantum transformer. *Neurocomputing* **2025**, *637*, 130107. [CrossRef]
19. Sethurajan, M.R.; Natarajan, K. Performance analysis of semantic veracity enhance (SVE) classifier for fake news detection and demystifying the online user behaviour in social media using sentiment analysis. *Soc. Netw. Anal. Min.* **2024**, *14*, 36. [CrossRef]
20. Bounaama, R.; Abderrahim, M.E.A. Classifying COVID-19 Related Tweets for Fake News Detection and Sentiment Analysis with BERT-based Models. *arXiv* **2023**, arXiv:2304.00636. [CrossRef]
21. Kula, S.; Choras, M.; Kozik, R.; Ksieniewicz, P.; Wozniak, M. Sentiment Analysis for Fake News Detection by Means of Neural Networks. In Proceedings of the ICCS 2020: International Conference on Computational Science, Amsterdam, The Netherlands, 3–5 June 2020; pp. 653–666.
22. Zhang, H.; Li, Z.; Liu, S.; Huang, T.; Ni, Z.; Zhang, J.; Lv, Z. Do Sentence-Level Sentiment Interactions Matter? Sentiment Mixed Heterogeneous Network for Fake News Detection. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 5090–5100. [CrossRef]
23. Kumari, R.; Ashok, N.; Ghosal, T.; Ekbal, A. A Multitask Learning Approach for Fake News Detection: Novelty, Emotion, and Sentiment Lend a Helping Hand. In Proceedings of the IJCNN 2021: International Joint Conference on Neural Networks, Shenzhen, China, 18–22 July 2021; pp. 1–8.

24. Jiang, S.; Guo, Z.; Ouyang, J. What makes sentiment signals work? Sentiment and stance multi-task learning for fake news detection. *Knowl. Based Syst.* **2024**, *303*, 112395. [CrossRef]
25. Zhou, X.; Wu, J.; Zafarani, R. SAFE: Similarity-Aware Multi-modal Fake News Detection. In Proceedings of the PAKDD 2020: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, 11–14 May 2020; pp. 354–367.
26. Wang, Z.; Huang, D.; Cui, J.; Zhang, X.; Ho, S.-B.; Cambria, E. A review of Chinese sentiment analysis: Subjects, methods, and trends. *Artif. Intell. Rev.* **2025**, *58*, 75. [CrossRef]
27. Sun, Q.H.; Gao, B. Emotion-enhanced Cross-modal Contrastive Learning for Fake News Detection. In Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security, Hangzhou, China, 21–23 February 2025; pp. 184–189.
28. Toughrai, Y.; Langlois, D.; Smaïli, K. Fake News Detection via Intermediate-Layer Emotional Representations. In Proceedings of the WWW '25: The ACM Web Conference 2025, Sydney, NSW, Australia, 28 April–2 May 2025; pp. 2680–2684. [CrossRef]
29. Tan, Z.H.; Zhang, T. Emotion-semantic interaction network for fake news detection: Perspectives on question and non-question comment semantics. *Inf. Process. Manag.* **2026**, *63*, 104391. [CrossRef]
30. Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Lu, T.; Shang, L. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In Proceedings of the WWW 2022: The Web Conference, Lyon, France, 25–29 April 2022; pp. 2897–2905.
31. Sun, M.; Zhang, X.; Ma, J.; Xie, S.; Liu, Y.; Yu, P.S. Inconsistent Matters: A Knowledge-Guided Dual-Consistency Network for Multi-Modal Rumor Detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12736–12749. [CrossRef]
32. Yu, H.; Wu, H.; Fang, X.; Li, M.; Zhang, H. SR-CIBN: Semantic relationship-based consistency and inconsistency balancing network for multimodal fake news detection. *Neurocomputing* **2025**, *635*, 129997. [CrossRef]
33. Wang, L.; Zhang, C.; Xu, H.; Xu, Y.; Xu, X.; Wang, S. Cross-modal Contrastive Learning for Multimodal Fake News Detection. In Proceedings of the ACM Multimedia 2023: 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 5696–5704.
34. Jiang, Y.; Wang, T.; Xu, X.; Wang, Y.; Song, X.; Maynard, D. Cross-modal augmentation for few-shot multimodal fake news detection. *Eng. Appl. Artif. Intell.* **2025**, *142*, 109931. [CrossRef]
35. Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; Ge, S. Bootstrapping Multi-View Representations for Fake News Detection. In Proceedings of the AAAI 2023: 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 5384–5392.
36. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019: Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the CVPR 2016: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; Tian, Q. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Trans. Multim.* **2017**, *19*, 598–608. [CrossRef]
39. Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; Kompatsiaris, Y. Detection and visualization of misleading content on Twitter. *Int. J. Multim. Inf. Retr.* **2018**, *7*, 71–86. [CrossRef]
40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the ICLR 2015: International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
41. Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* **2021**, *58*, 102610. [CrossRef]
42. Jing, J.; Wu, H.; Sun, J.; Fang, X.; Zhang, H. Multimodal fake news detection via progressive fusion networks. *Inf. Process. Manag.* **2023**, 60. [CrossRef]
43. Peng, L.; Jian, S.; Kan, Z.; Qiao, L.; Li, D.S. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Inf. Process. Manag.* **2024**, *61*, 103564. [CrossRef]
44. Yang, H.; Zhang, J.; Zhang, L.; Cheng, X.; Hu, Z. MRAN: Multimodal relationship-aware attention network for fake news detection. *Comput. Stand. Interfaces* **2024**, *89*, 103822. [CrossRef]