A decoupled approach for multimodal fake news detection incorporating entity knowledge enhancement and multi-scale adaptive fusion

Ailin Li, Wendi Zhu*, Zhihang Wang, Jizong Wang, Wei Sheng

Department of Artificial Intelligence, Gansu University of Political Science and Law, Lanzhou, Gansu 730030, China

Keywords: Knowledge enhancement; Multi-scale features; Dynamic dual-threshold fusion; Decoupled architecture; Multimodal fake news detection.

Citation: Li A.L., Zhu W.D., Wang Z.H., et al.: A decoupled approach for multimodal fake news detection incorporating entity knowledge enhancement and multi-scale adaptive fusion. Data Intelligence, 2026, Vol. 8, Art. No.: 20251349. DOI: https://doi.org/10.3724/2096-7004.di.2025.1349

ABSTRACT

Multimodal fake news spreads through coordinated dissemination of textual and visual content, posing threats to public opinion and social safety. Current approaches exhibit three major limitations: inadequate cross-modal semantic association processing, insufficient entity fact verification, and inflexible fusion strategies. First, semantic mappings between textual and visual modalities are inherently complex, hindering accurate cross-modal consistency detection. Second, fake news frequently propagates through fabricated entities, yet current methods lack robust entity verification capabilities. Third, fusion strategies employ static weights and cannot adapt to varying cross-modal consistency, thereby limiting performance in complex scenarios. To address these limitations, this paper proposes EKAMF-Net, a decoupled model that integrates entity knowledge enhancement with multi-scale adaptive fusion. The model employs a decoupled architecture with three core components: (1) multi-scale feature extraction using RoBERTa (Robustly Optimized BERT Pretraining Approach) and CLIP (Contrastive Language-Image Pre-training) ViT-L/14 for textual and visual data, (2) a Wikipedia-based entity knowledge enhancement module that integrates factual information through attention mechanisms, and (3) a dynamic dual-threshold fusion strategy that categorizes cross-modal consistency into three states—strong consistency, weak consistency, and significant conflict—for adaptive weight adjustment. Experimental results demonstrate that EKAMF-Net achieves macro-averaged F1 score of 93.0% and 91.8% on Twitter and Weibo datasets, respectively, representing an average improvement of 2 percentage points over the best baseline.

*Corresponding author: Wendi Zhu (E-mail: zhuwendi@stu.gsupl.edu.cn; ORCID: 0009-0002-5128-0682).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Additionally, the model demonstrates improvements of 5.7%–6.3% in implicit entity conflict detection, establishing a robust solution for multimodal fake news detection.

1. INTRODUCTION

The rapid development of internet technology and the widespread adoption of social media have facilitated the rapid and widespread dissemination of information. However, this landscape has also been exploited for the proliferation of multimodal fake news, which poses significant societal risks [1-2]. Multimodal fake news combines textual and visual content to mislead public perception, manipulate opinion, and potentially trigger social unrest [3]. For instance, during public health emergencies, false medical information and altered images can cause public misunderstanding of prevention measures, disrupting epidemic control efforts. Similarly, during elections, biased content and edited images may influence voter judgments, undermining electoral fairness [4]. Therefore, research on multimodal fake news detection holds significant theoretical and practical value.

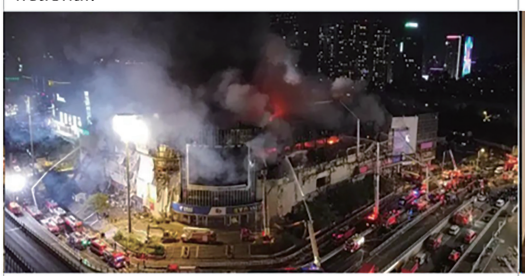
Although existing multimodal fake news detection methods have achieved considerable progress, they face three major challenges [5-6]. First, cross-modal semantic associations are inherently complex. The semantic mapping between textual and visual content involves ambiguity and diversity rather than simple one-to-one correspondence. The same image may convey different meanings across textual contexts, while identical text may correspond to various visual representations. These complex associations hinder accurate detection of cross-modal consistency [7]. Second, multimodal fake news has proliferated extensively across social media platforms, with significant portions misleading the public through entity falsification [8]. Figure 1 illustrates typical entity-based fake news cases: false disaster reports fabricate key information, including locations and timestamps to create panic, while fabricated accident scenes provide incorrect geographical data and casualty figures. Such fake news incorporates specific entities (place names, timestamps, casualty figures) within text, accompanied by seemingly authentic images, thereby enhancing deceptive potential. Traditional multimodal detection methods focus primarily on semantic consistency between textual and visual content but prove ineffective against entity-level fabrication. This occurs because textual descriptions and visual content may appear superficially consistent, requiring entity fact-checking to reveal their false nature [9]. Third, fusion strategies lack dynamic adaptability. Current multimodal fusion methods employ fixed weights or thresholds and cannot adapt to varying cross-modal consistency or data characteristics, thereby limiting detection performance in complex scenarios [10].



"On January 23, 2024, a landslide disaster occurred in a certain area of Yunnan Province, resulting in the deaths of eight people. Rescue efforts are currently underway at the scene." There are no records of landslide disasters in Yunnan Province during this time period, and the incident is entirely fictional.



"On June 9,2024, a 5.0-magnitude earthquake struck Xide County in Liangshan Yi Autonomous Prefecture, Sichuan Province. The epicenter and surrounding areas suffered severe damage to buildings, resulting in multiple casualties." The actual earthquake occurred in Muli County, not Xide County.



"On the evening of January 10, 2024, a violent explosion occurred in a commercial district in Xi'an, causing flames to shoot into the sky and destroying several vehicles. The authorities have launched an investigation." There were no explosions in Xi'an during this period. All entities are false.



"A residential building in Nan'an, Chongqing caught fire, killing 19 people." This text exaggerates the disaster by falsifying the number of casualties and is a typical example of false information.

Figure 1. Typical case of entity-based fake news.

To address these challenges, this paper proposes EKAMF-Net, a decoupled model employing entity knowledge enhancement and multi-scale adaptive fusion for precise multimodal fake news detection. The model introduces three key innovations: First, the model adopts decoupled multi-scale feature extraction. Feature extraction processes for textual and visual data are decoupled, with separate multi-scale modules designed for each modality [11]. For textual data, RoBERTa extracts features at word, sentence, and paragraph levels to capture comprehensive semantic information. For visual data, CLIP-ViT-L/14 extracts pixel, region, and scene-level features to capture comprehensive visual information [12]. This multi-scale approach captures information at different granularities, providing rich features for subsequent fusion and detection. Second, the model incorporates a Wikipedia-based entity knowledge enhancement module. This module associates entities extracted from textual and visual content with Wikipedia, obtaining factual information, including entity attributes and relationships. These knowledge features are then integrated into multi-scale features through attention mechanisms [13]. This approach enhances detection capabilities

from semantic-level to fact-level analysis, improving entity-based fake news identification. Third, the model implements a dynamic dual-threshold adaptive fusion strategy. Based on cross-modal feature consistency, the system dynamically learns dual thresholds to classify relationships between textual and visual content into three states: strong consistency, weak consistency, and significant conflict. For each state, fusion weights are adaptively adjusted to achieve effective multi-scale feature fusion and improve adaptability in complex scenarios.

To validate EKAMF-Net effectiveness, extensive experiments were conducted on two publicly available multimodal fake news datasets: Twitter and Weibo [15-16]. Experimental results demonstrate that EKAMF-Net outperforms existing baseline models in evaluation metrics including F1(macro-F1) score and accuracy, with particular advantages in detecting implicit entity conflicts. Additionally, ablation studies and comparative analysis validate the effectiveness and rationale of each model component.

2. RELATED WORK

2.1 Research Progress in Multimodal Fake News Detection

Multimodal fake news detection has emerged as a research focus in artificial intelligence and information security, with the core objective of leveraging multiple modalities, including text and images for accurate fake news identification [17]. Early multimodal fake news detection methods primarily relied on traditional machine learning algorithms with manually crafted features. For instance, Kumar et al. (2016) [18] proposed a multimodal fake news detection method based on sentiment analysis and visual features. This method extracts sentiment features from text and color/texture features from images, then employs support vector machines (SVM) for classification. However, these manually crafted feature methods often rely on domain knowledge, exhibit poor generalization capabilities, and struggle to address complex and evolving forms of fake news.

With advances in deep learning technology, deep learning-based multimodal fake news detection methods have gradually become mainstream [19]. These methods leverage neural networks to automatically learn feature representations from text and images, which has led to improved detection performance. For instance, Khattar et al. (2019) [14] proposed a multimodal fusion model based on CNN (convolutional neural network) and RNN (recurrent neural network). This model employs CNN to extract visual features and RNN to extract textual features, which are then fused through concatenation to achieve multimodal representation learning. Additionally, advancements in image steganography analysis and anti-counterfeiting technologies provide technical support for multimodal fake news detection. Liao et al. (2022) [20] proposed an adaptive payload allocation method based on image texture features, which enhances security in multi-image steganography through texture complexity analysis. Their feature extraction approach offers valuable insights for verifying the authenticity of multimodal information. Regarding video authenticity detection, Liao et al. (2023) [21] proposed a detection framework based on facial muscle movements for compressed Deepfake videos circulating on social networks. This method

identifies forgery traces in compressed videos by capturing dynamic facial features, offering a novel approach for authenticating low-quality multimodal content on social media.

Recently, pre-trained model-based multimodal fake news detection methods have achieved significant progress [22]. Pre-trained models learn rich general features through large-scale dataset pre-training, thereby demonstrating outstanding performance in downstream tasks. For instance, the CLIP model proposed by Radford et al. (2021) [23] achieves cross-modal semantic alignment between textual and visual content through pre-training on large-scale image-text pairs, providing a powerful tool for multimodal fake news detection. Choi et al. (2025) [24] proposed the CroMe model, which utilizes BLIP2-OPT as the encoder to capture text, image, and combined image-text representations, and employs a cross-modal tri-transformer for feature fusion and a proxy anchor method for metric learning to enhance intra-modality relationship capture. However, the CroMe model fails to account for hierarchical differences in multi-scale features: it only extracts single-scale global features from both modalities, ignoring the fact that fine-grained visual tampering often occurs at the pixel or region level, while entity conflicts are usually manifested at the word or sentence level. This limitation results in insufficient detection capabilities for fine-grained visual tampering and fine-grained entity conflicts.

Furthermore, some studies focus on optimizing multimodal fusion strategies [5]. For instance, Li et al. (2023) [25] proposed the CFFN model, which calculates cross-modal correlations through fine-grained fragment matching and identifies consistent and conflicting regions based on fixed thresholds to achieve multimodal feature fusion. However, this model employs fixed thresholds, thereby limiting generalization capability and causing significant performance variations across different datasets and scenarios.

2.2 Current Status of Research on Entity Knowledge Enhancement Technology

Entity knowledge enhancement techniques improve model understanding of entities and factual judgment capabilities by integrating entity information from external knowledge bases. These techniques have been extensively applied across natural language processing and computer vision [26]. In multimodal fake news detection, entity knowledge enhancement techniques help models identify entity authenticity and consistency, thereby improving detection performance. Early entity knowledge enhancement methods were primarily grounded in knowledge graphs, embedding entities and relationships into model feature spaces [27]. For instance, Zhang et al. (2019) [28] proposed the ERNIE(Enhanced Representation Through Knowledge Integration) model, which enhances textual semantic understanding by integrating knowledge graph entity information into BERT. However, this method primarily targets textual modalities and has not been extended to visual modalities, precluding multimodal entity knowledge fusion. Recently, with advances in cross-modal technologies, researchers have begun exploring multimodal entity knowledge enhancement methods [4]. Wang et al. (2024) [29] proposed the KAT framework, which employs entity linking technology to associate entities in textual and visual content with knowledge graphs, achieving entity knowledge fusion in multimodal features. This framework has demonstrated effectiveness in entity fact verification tasks but relies on knowledge graph completeness and exhibits limited coverage of emerging and long-tail entities, thereby limiting its applicability in dynamic news scenarios. Wikipedia,

as the world's largest open knowledge repository, contains extensive entity information with real-time update capabilities, making it highly suitable for entity knowledge enhancement in multimodal fake news detection [30]. Several studies have begun utilizing Wikipedia for entity knowledge enhancement. For instance, Ferragina et al. (2012) [8] proposed the TagMe tool, which automatically identifies entities in text and links them to Wikipedia, facilitating textual entity knowledge enhancement. However, the effective integration of Wikipedia entity knowledge into multimodal features for cross-modal entity knowledge enhancement remains a critical challenge.

2.3 Research on Multi-Scale Feature Fusion and Adaptive Fusion Strategies

Multi-scale feature fusion technology integrates feature information across different scales, comprehensively capturing both detailed and global data information, and has been extensively applied in computer vision and natural language processing [31]. In multimodal fake news detection, both textual and visual data possess multi-scale features, and integrating features across different scales enhances the model's capability to detect complex fake news. In computer vision, multi-scale feature fusion technology has achieved considerable maturity [32]. For instance, the ResNet model proposed by He et al. (2016) [33] achieves feature fusion across different layers through residual connections, thereby improving image classification performance. The ConvNeXt model proposed by Liu et al. (2022) [34] extends ResNet by introducing multi-stage convolutions to extract multi-scale features from images, further enhancing image understanding capabilities. These methods provide valuable insights for image multi-scale feature fusion in multimodal fake news detection. In natural language processing, multi-scale feature fusion primarily extracts lexical, sentence-level, and paragraph-level features from textual data through neural networks at different levels [35]. For instance, the BERT model proposed by Devlin et al. (2019) [11] ploys multi-layer Transformers to extract deep semantic features from text, where features at different levels reflect semantic information at varying granularities. The RoBERTa model (Liu et al., 2019) [36] optimizes BERT by adjusting training strategies and data volume, further enhancing textual feature extraction capabilities. Adaptive fusion strategies dynamically adjust fusion weights based on data characteristics and task requirements, thereby improving multimodal fusion effectiveness [37]. In time series prediction, Wu et al. (2025) [13] proposed the TimeMixer model, which decouples past information and future predictions in multi-scale time series and utilizes a multi-scale hybrid architecture to achieve adaptive time series fusion, thereby improving prediction performance. This adaptive fusion approach provides insights for designing fusion strategies in multimodal fake news detection but requires targeted adjustments based on multimodal data characteristics. Liao et al. (2024) [38] proposed a grayscale deviation simulation mechanism in their study on screenshot watermarking. By adaptively modeling distortions during screen display and camera capture processes, they constructed a more realistic noise layer. Liao et al. (2024) [39] further proposed a cascaded network based on the wavelet domain, implementing adaptive feature fusion and recovery mechanisms within the wavelet domain. Although applied in different domains, the design philosophy of these adaptive strategies—dynamically adjusting processing strategies based on data characteristics—shares common ground with the dynamic dual-threshold fusion strategy proposed in this paper. This provides cross-domain methodological references for the adaptive fusion of multimodal features.

In summary, existing multimodal fake news detection methods exhibit limitations in multi-scale feature extraction, entity knowledge enhancement, and adaptive fusion strategies. The proposed EKAMF-Net model integrates multi-scale feature extraction, Wikipedia-based entity knowledge enhancement, and dynamic adaptive fusion strategies to address existing method limitations and enhance multimodal fake news detection performance.

3. MODEL STRUCTURE (EKAMF-NET)

3.1 Overall architecture

The EKAMF-Net model employs a decoupled architecture composed of four core sub-modules: a multi-scale feature extraction module, an entity knowledge enhancement module, a dynamic threshold learning module, and an adaptive fusion-classification module. The overall framework is depicted in Figure 2. Specifically, the dynamic threshold learning module and the adaptive fusion-classification module work collaboratively to achieve adaptive fusion, with the former generating sample-specific thresholds and the latter utilizing these thresholds to refine the classification decision.

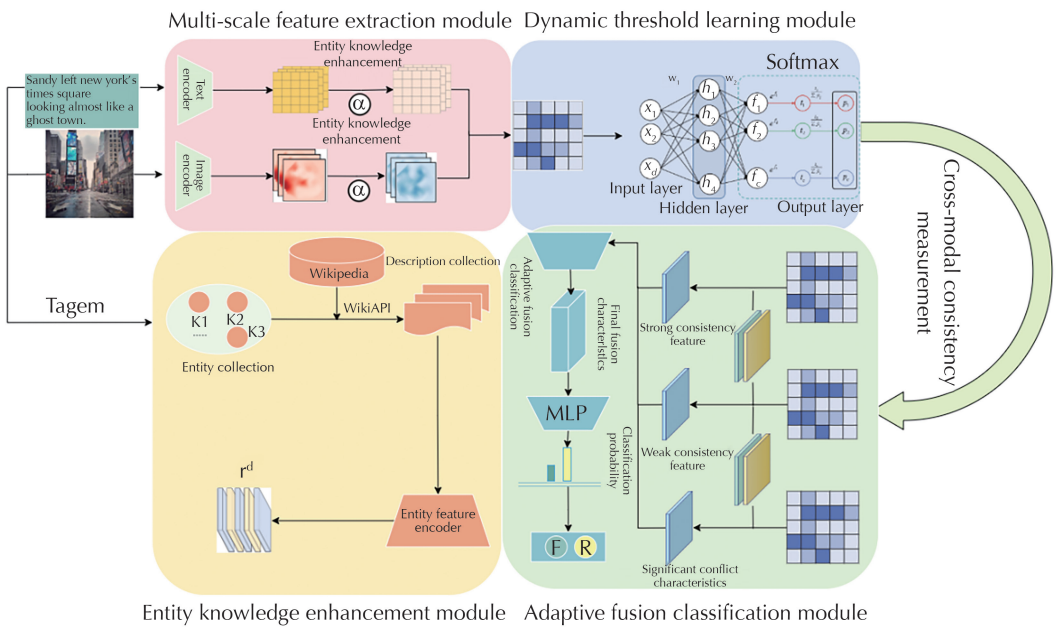


Figure 2. EKAMF-Net Overall Architecture Diagram.
(Note: Pink represents the multi-scale feature extraction module, green represents the multi-scale feature adaptive fusion module, blue represents the dynamic threshold learning module, and yellow represents the entity knowledge enhancement module. The @ symbol indicates attention mechanisms, and arrows indicate data flow directions.)

The multi-scale feature extraction module extracts multi-scale features from textual and visual data, providing foundational features for subsequent processing. The entity knowledge enhancement module associates entities in textual and visual content with the Wikipedia knowledge base to generate entity knowledge features, which are subsequently integrated into multi-scale features to enhance factual accuracy. The dynamic threshold and adaptive fusion modules dynamically adjust fusion weights based on cross-modal feature consistency, achieving effective multi-scale feature fusion and ultimately outputting fake news detection results.

3.2 Multi-scale feature extraction module

The multi-scale feature extraction module employs a decoupled design, performing separate multi-scale feature extraction for textual and visual data to comprehensively capture semantic and visual information.

3.2.1 Multi-scale feature extraction from text

It has been widely demonstrated that different layers of pre-trained Transformer models capture semantic information at varying granularities. Tenney et al. (2019) revealed through edge detection tasks that BERT's shallow layers primarily encode syntactic information such as phrase structure, while its middle layers capture dependency syntax and anaphoric relations, and its deep layers focus on high-level semantics like semantic roles [40]. Jawahar et al. (2019) further validated this hierarchical feature encoding pattern through multiple probe tasks [41]. RoBERTa enhances model performance by optimizing the pre-training strategy while retaining the same Transformer layer structure as BERT. Based on these studies, this paper extracts lexical-level features from RoBERTa's shallow layers, sentence-level features from its middle layers, and discourse-level features from its deep layers.

Textual multi-scale feature extraction employs the RoBERTa-base model as the base encoder, which comprises 12 Transformer layers and extracts multi-scale features from different textual levels. Specifically, the textual multi-scale feature extraction process proceeds as follows:

Word-level features primarily capture the semantic information of words within the text, including entity names and emotional words. Word-level features are extracted from layers 2–3 of the RoBERTa model, as features from these layers better reflect the local semantic information of words. Employing a Gaussian Error Linear Unit (GELU) activation function to enhance nonlinear representation capability. The word-level feature calculation formula is as follows:

$$T_{low} = \mathrm{GELU}(\mathrm{RoBERTa}_{2-3}(x_{text})) \tag{1}$$

Here, T_{low} symbolizes word-level features that provide a fine-grained semantic basis for entity recognition and local consistency analysis. x_{text} denotes the input text sequence, $\mathrm{RoBERTa}_{2-3}(.)$ represents the output of layers 2–3 of the RoBERTa model. Word-level features have a dimension of 768.

Sentence-level features primarily capture logical relationships and semantic structures within textual sentences, including chronological event ordering and causal relationships. Sentence-level features are extracted from layers 6–7 of the RoBERTa model, as features from these layers better reflect intermediate-level semantic information of sentences. The sentence-level feature calculation formula is as follows:

$$T_{mid} = \text{GELU}(\text{RoBERTa}_{6-7}(x_{text})) \tag{2}$$

Here, T_{mid} represents sentence-level features that facilitate the modeling of intermediate semantic structures for cross-modal alignment. $\text{RoBERTa}_{6-7}(.)$ refers to the output of layers 6–7 of the RoBERTa model, with sentence-level features having a dimension of 768.

Paragraph-level features primarily capture the overall theme and stance of textual content, reflecting high-level semantic information. Paragraph-level features are extracted from layers 11–12 of the RoBERTa model through the [CLS] token. The paragraph-level feature calculation formula is as follows:

$$T_{high} = \text{GELU}(\text{RoBERTa}_{11-12}(x_{text})[\text{CLS}]) \tag{3}$$

Here, T_{high} denotes paragraph-level features that encapsulate global semantics, crucial for high-level consistency evaluation. $\text{RoBERTa}_{11-12}(.)$ [CLS] represents features corresponding to the [CLS] token in the output of layers 11–12 of the RoBERTa model, with paragraph-level features having a dimension of 768.

3.2.2 Multi-scale feature extraction from images

It has been observed that different layers of the Vision Transformer learn distinct levels of visual representation. Raghu et al. (2021) revealed through representational similarity analysis and attention map visualization that shallow ViT representations primarily focus on local textures and edge features. As layers deepen, they progressively develop an understanding of object components, while deeper layers integrate global scene semantics [42]. Through large-scale visual-language alignment training, CLIP's ViT encoder similarly exhibits a hierarchical structure progressing from fine-grained visual features to high-level semantic representations. Based on this, this paper extracts pixel-level features from shallow layers, region-level features from intermediate layers, and scene-level features from deep layers of CLIP-ViT-L/14.

For image multi-scale feature extraction, we employ the CLIP-ViT-L/14 model as the base encoder. This model consists of 14 × 14 patches and 24 Transformer layers, facilitating multi-scale feature extraction at different hierarchical levels. The extraction process operates at three distinct levels:

Pixel-level features capture fine-grained visual information, including texture and color patterns, which reflect fine-grained image characteristics essential for detecting tampering artifacts. These features are extracted from layers 2–4 of the CLIP-ViT-L/14 model, which effectively encode fine-grained visual representations. The calculation formula is as follows:

$$I_{low} = \text{GELU}(\text{CLIP-ViT-L/14}_{2-4}(x_{img})) \tag{4}$$

In this context, I_{low} denotes pixel-level features that serve as the visual basis for detecting fine-grained anomalies. x_{img} represents the input image data, while CLIP-ViT-L/$14_{2-4}(.)$ denotes the output of layers 2–4 of the CLIP-ViT-L/14 model, with pixel-level features having a dimension of 768.

Regional-level features primarily capture intermediate-level visual information such as the shape, position, and interrelationships of objects in an image, enabling the identification of key entities and regions within the image. Regional-level features are extracted from layers 10–12 of the CLIP-ViT-L/14 model, as these layers effectively reflect the intermediate-level visual information in images. To enhance the ability of regional-level features to distinguish entities, the SiLU(Sigmoid Linear Unit) activation function is applied. The formula for calculating regional-level features is as follows:

$$I_{mid} = \text{SiLU}(\text{CLIP-ViT-L/}14_{10-12}(x_{img})) \tag{5}$$

Among them, I_{mid} represents region-level features that enhance entity discrimination for multimodal alignment. CLIP-ViT-L/$14_{10-12}(.)$ represents the output of layers 10–12 of the CLIP-ViT-L/14 model, with a dimension of 768 for region-level features.

Scene-level features primarily capture high-level visual information such as the overall environment and atmosphere of an image, reflecting its overall semantic meaning. Scene-level features are extracted from layers 22–24 of the CLIP-ViT-L/14 model and obtained through the model's [CLS] token. The formula for calculating scene-level features is as follows:

$$I_{high} = \text{GELU}(\text{CLIP-ViT-L/}14_{22-24}(x_{img})[\text{CLS}]) \tag{6}$$

Among them, Ihigh symbolizes scene-level features that enable global semantic integration for consistency evaluation. CLIP-ViT-L/$14_{22-24}(.)$ [CLS] represents the features corresponding to the [CLS] token in the output of layers 22–24 of the CLIP-ViT-L/14 model, and the dimension of scene-level features is 768.

3.3 Entity Knowledge Enhancement Module

To address entity authenticity verification in multimodal fake news, we designed an entity knowledge enhancement module that validates entities using external knowledge bases. Figure 3 illustrates the core workflow: entity extraction from text, knowledge retrieval from Wikipedia, and feature integration to enhance entity-level fake news detection. This module associates textual and visual entities with Wikipedia, retrieves factual information, and injects entity knowledge features into multi-scale representations to enhance entity authenticity assessment. The module comprises three submodules: entity recognition and linking, entity knowledge feature generation, and entity knowledge feature injection.

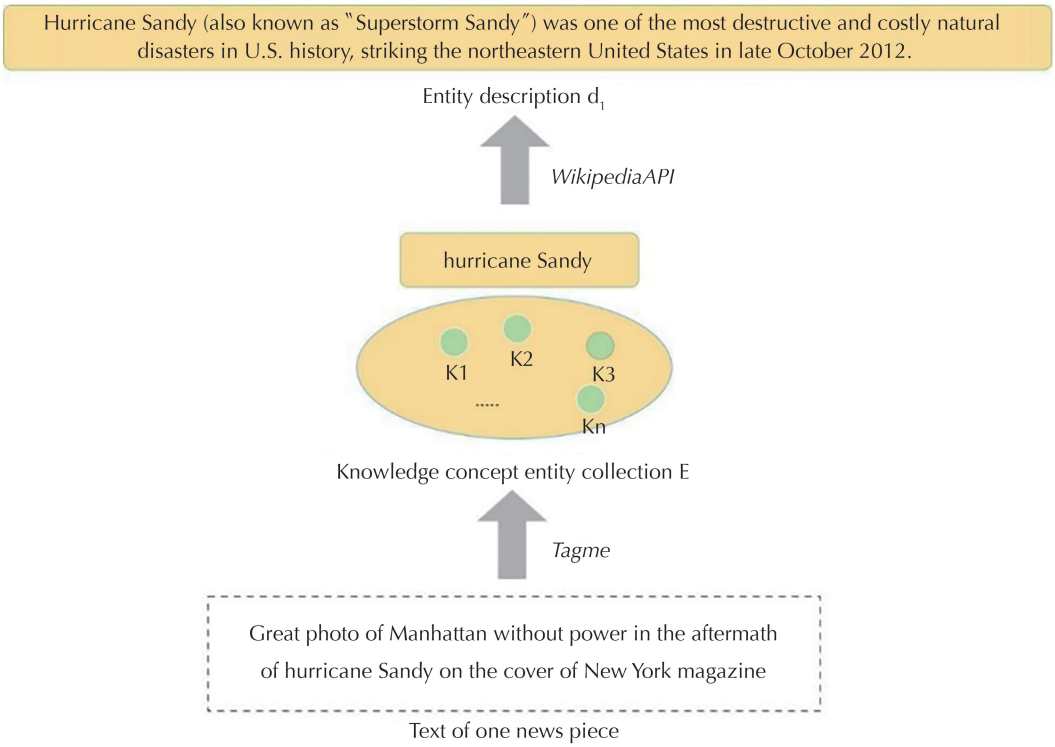


Figure 3. Core workflow of entity knowledge enhancement.

3.3.1 Entity recognition and linking

Entity recognition and linking identify entities from text and images, and associate them with Wikipedia entries, providing a foundation for knowledge feature generation.

For textual data, the TagMe entity linking tool is applied to the original text for entity recognition and linking. The TagMe tool automatically identifies textual entities and calculates confidence scores for each identified entity. A confidence threshold of 0.6 is applied. entities with confidence scores ≥ 0.6 are retained and aligned with the word-level features Tlow extracted by RoBERTa for subsequent knowledge feature generation.

For visual data, region-level features I_{mid} are extracted from CLIP-ViT-L/14 to generate region proposals. Textual descriptions for these proposals are then generated, and the text-based entity recognition tool TagMe is used to process the descriptions, obtaining corresponding Wikipedia entity descriptions.

3.3.2 Generation of entity knowledge features

Entity knowledge feature generation converts Wikipedia entity descriptions into vector representations for fusion with multi-scale features. RoBERTa encodes entity descriptions to generate knowledge features.

Specifically, entity descriptions are input to RoBERTa, and features from the [CLS] token in the 12th layer are extracted as entity knowledge features E. The calculation formula is:

$$E = \text{RoBERTa}(d_{wiki})[\text{CLS}] \tag{7}$$

Here, d_{wiki} represents Wikipedia entity descriptions, and entity knowledge features E have dimension 768.

3.3.3 Entity knowledge feature injection

Entity knowledge feature injection integrates generated entity knowledge features E into multi-scale textual and visual features to enhance their factuality. This integration employs attention mechanisms.

As shown in Figure 4, the entity knowledge feature injection process comprises the following key steps. First, extract entity information from image region-level features Imid and text lexical-level features Tlow; then obtain corresponding Wikipedia knowledge descriptions through entity recognition and linking; subsequently encode the knowledge descriptions into entity knowledge features E; finally, inject the entity knowledge features into text sentence-level features Tmid and image region-level features Imid respectively via an attention mechanism to generate enhanced feature representations.

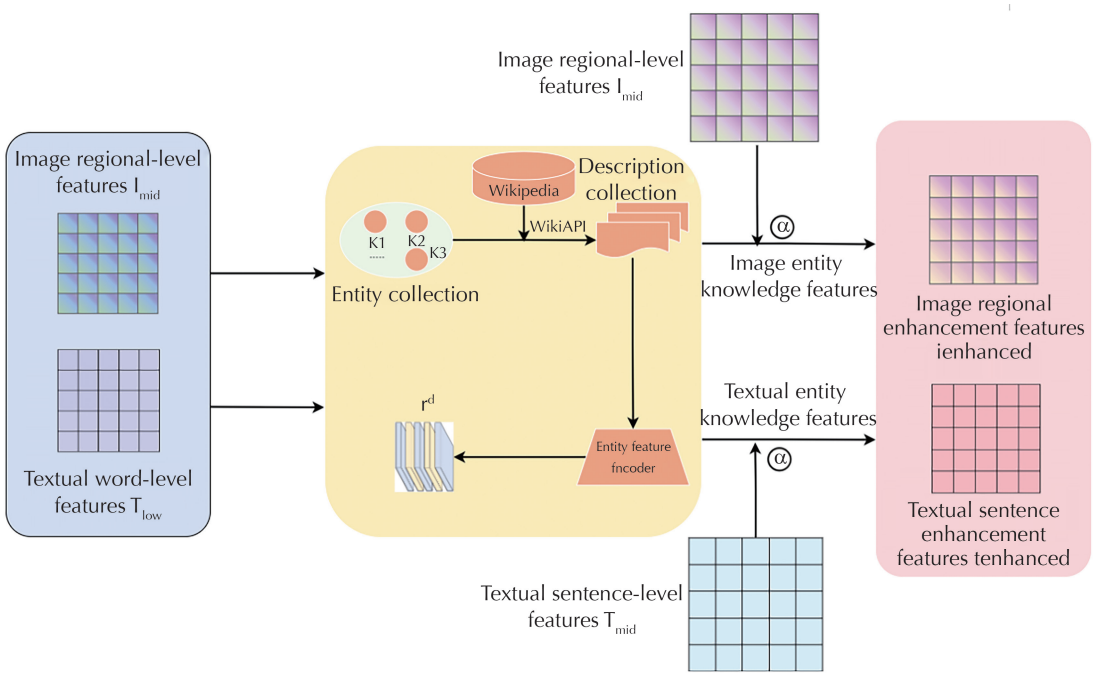


Figure 4. Data Flow and Feature Fusion Process of the Entity Knowledge Augmentation Module.

Sentence-level textual features T_{mid} serve as query vectors (Q), while entity knowledge features (E) serve as both key vectors (K) and value vectors (V). Attention mechanisms calculate correlation weights between textual and entity knowledge features, which are then weighted and injected into sentence-level textual features to obtain enhanced features $T_{enhanced}$. The calculation formulas are:

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{8}$$

$$T_{enhanced} = T_{mid} + \text{Attention}(T_{mid},\ E,\ E) \tag{9}$$

Among them, d_k is the dimension of the key vector, where $d_k = 768$.

Similarly, for visual features, regional features I_{mid} serve as query vectors (Q), while entity knowledge features (E) serve as key and value vectors (K, V). Correlation weights between visual and entity knowledge features are calculated and applied to obtain enhanced regional features $I_{enhanced}$. The calculation formula is:

$$I_{enhanced} = I_{mid} + \text{Attention}(I_{mid},\ E,\ E) \tag{10}$$

Simultaneously, to quantify entity conflict, we calculate the cosine similarity between image region-level features I_{mid} and entity knowledge features (E) to generate entity conflict scores (Sent). The calculation formula is:

$$S_{ent} = 1 - \cos(I_{mid},\ E) \tag{11}$$

Among them, $\cos(\cdot, \cdot)$ represents the cosine similarity function, and S_{ent} takes values in the range [0, 2]. The larger the value, the more significant the conflict between the image entity and the Wikipedia entity.

3.4 Dynamic threshold learning module

The dynamic threshold learning module learns two dynamic thresholds based on cross-modal feature consistency strength to classify relationships into different consistency states. This module comprises two submodules: cross-modal consistency measurement and dynamic threshold learning.

3.4.1 Cross-modal consistency measurement

Cross-modal consistency measurement quantifies semantic consistency between textual and visual content, providing a basis for threshold learning and feature fusion. Considering high-level semantic features and entity conflict scores, we calculate the comprehensive cross-modal consistency score S_{total}.

We calculate cosine similarity between paragraph-level textual features T_{high} and scene-level visual features I_{high} as the semantic consistency score S_{sem}:

$$S_{sem} = \cos(T_{high},\ I_{high}) \tag{12}$$

Among them, the value range of S_{sem} is [–1, 1], and the larger the value, the stronger the semantic consistency between the text and the image.

The semantic consistency score S_{sem} and entity conflict score S_{ent} are weighted and fused to obtain the comprehensive consistency score S_{total}. Through experimental validation, the semantic consistency weight is set to 0.7 and the entity conflict weight to 0.3:

$$S_{total} = 0.7 \times \frac{S_{sem} + 1}{2} + 0.3 \times \frac{2 - S_{ent}}{2} \tag{13}$$

Among these, S_{sem} is normalized to the range [0, 1], S_{ent} is normalized to the range [0, 1], and S_{total} is normalized to the range [0, 1], where a higher value indicates stronger overall consistency between the text and image.

3.4.2 Dynamic threshold learning

The purpose of dynamic threshold learning is to learn two dynamic thresholds τ_1 and τ_2 ($\tau_1 < \tau_2$) based on the distribution characteristics of the cross-modal consistency score S_{total}, and to classify the cross-modal relationship between text and images into three states: "strong consistency", "weak consistency", and "significant conflict".

A two-layer multi-layer perceptron (MLP) is used to learn dynamic thresholds. The mean $\mu(S_{total})$ and variance $\sigma(S_{total})$ of the cross-modal synthesis consistency score S_{total} are used as inputs to the MLP. After processing by the two-layer neural network, two dynamic thresholds τ_1 and τ_2 are output via the Sigmoid activation function. The calculation formula is as follows:

$$[\tau_1, \tau_2] = \text{Sigmoid}(\text{MLP}(\mu(S_{total}), \sigma(S_{total}))) \tag{14}$$

In this model, the first layer of the MLP contains 256 neurons using the GELU activation function; the second layer contains 2 neurons, and the output is processed by the Sigmoid activation function to restrict the threshold to the [0, 1] interval. Based on the learned dynamic thresholds, cross-modal relationships are categorized into three states: when $S_{total} \geq \tau_2$, the text and image are in a "strong consistency" state; when $\tau_1 \leq S_{total} \leq \tau_2$, the text and image are in a "weak consistency" state; and when $S_{total} < \tau_1$, the text and image are in a "significant conflict" state.

3.5 Adaptive Fusion Classification Module

The adaptive fusion classification module learns cross-modal relationship states from the dynamic threshold learning module and applies different fusion strategies to combine multi-scale textual and visual features. It then outputs fake news detection results through a classifier. This module comprises two submodules: multi-scale feature fusion and classification output.

3.5.1 Multi-scale feature adaptive fusion

Multi-scale feature fusion: Based on cross-modal relationship states, different fusion strategies are applied to combine multi-scale textual and visual features, obtaining final fusion features for fake news detection through a classifier.

When textual and visual content are in a "strongly consistent" state, they exhibit high consistency at both the semantic and entity levels. In this case, fusion focuses on high-level textual and visual features. Enhanced text sentence-level features $T_{enhanced}$, text paragraph-level features T_{high}, enhanced image region-level features $I_{enhanced}$, and image scene-level features I_{high} are concatenated and passed through a fully connected layer to obtain fused features F. The calculation formula is:

$$F = FC(Concat(T_{enhanced}, T_{high}, I_{enhanced}, I_{high})) \tag{15}$$

Among them, Concat(.) represents a feature concatenation operation, FC(.) represents a fully connected layer, and the dimension of feature fusion F is 1024.

When textual and visual content are in a "weak consistency" state, they exhibit partial semantic consistency but contain entity-level conflicts or uncertainties. In this case, entity knowledge features require enhanced influence. Enhanced text sentence-level features $T_{enhanced}$, text paragraph-level features T_{high}, enhanced image region-level features $I_{enhanced}$, image scene-level features I_{high}, and entity knowledge features E are concatenated and passed through a fully connected layer to obtain fused features F. The calculation formula is:

$$F = FC(Concat(T_{enhanced}, T_{high}, I_{enhanced}, I_{high}, E)) \tag{16}$$

When textual and visual content are in a "significant conflict" state, they exhibit clear inconsistency. In this case, the system determines which modality is more reliable based on entity knowledge features and emphasizes that modality's features. The system calculates similarity between enhanced textual features and entity knowledge features, as well as between enhanced visual features and entity knowledge features. Modality reliability is determined based on similarity magnitude. Features from the more reliable modality are weighted and fused with features from the other modality to obtain fused features F. The calculation formulas are:

$$sim_t = \cos(T_{enhanced}, E) \tag{17}$$

$$sim_i = \cos(I_{enhanced}, E) \tag{18}$$

$$w_t = \frac{\exp(sim_t)}{\exp(sim_t) + \exp(sim_t)} \tag{19}$$

$$w_i = 1 - w_t \tag{20}$$

$$F = w_t \times Concat(T_{enhanced}, T_{high}) + w_i \times Concat(I_{enhanced}, I_{high}) \tag{21}$$

Here, sim_t and sim_t represent cosine similarities between textual and visual enhanced features with entity knowledge features, respectively, while w_t and w_i represent fusion weights for textual and visual features, respectively.

3.5.2 Classification and loss function

Fusion feature F is input to a fully connected layer, and fake news prediction probability P is output through the Sigmoid activation function. The calculation formula is:

$$p = \text{Sigmoid}(FC(F)) \tag{22}$$

EKAMF-Net training employs a multi-task loss function comprising four components: classification loss, contrast loss, entity alignment loss, and threshold regularization loss. The calculation formula is:

$$\mathcal{L} = \mathcal{L}_{cls} + 0.2\mathcal{L}_{cons} + 0.25\mathcal{L}_{ent} + 0.1\mathcal{L}_{thresh} \tag{23}$$

The weights of each loss term in this formula will be verified in Section 4.5 through parameter sensitivity analysis to explain why these values are selected.

A binary cross-entropy loss function is used to optimize the classification performance of the model. The calculation formula is as follows:

$$\mathcal{L}_{cls} = -\left(y\log(p) + (1-y)\log(1-p)\right) \tag{24}$$

In this case, y represents the true label of the sample (1 indicates fake new, 0 indicates true information), and p represents the predicted probability of the model.

Used to narrow the distance between the text and image features of real samples and widen the distance between the textual and visual features of fake samples. The calculation formula is as follows:

$$\mathcal{L}_{cons} = y \times \| T_{enhanced} - I_{enhanced} \|_2^2 + (1-y) \times \max(0,\ m - \| T_{enhanced} - I_{enhanced} \|_2^2) \tag{25}$$

Among them, m is the marginal parameter, set to 1.0.

Used to penalize the distance between image region-level features and entity knowledge features, forcing image entities to align with Wikipedia entities. The calculation formula is as follows:

$$\mathcal{L}_{ent} = \| I_{mid} - E \|_2^2 \tag{26}$$

Used to constrain the distance between two dynamic thresholds, ensuring that the thresholds can effectively distinguish between different cross-modal relationship states. The calculation formula is as follows:

$$\mathcal{L}_{thresh} = (\tau_2 - \tau_1 - 0.4)^2 \tag{27}$$

Among them, the ideal distance between the two dynamic thresholds is set to 0.4, which is determined through preliminary experiments, semantic rationality analysis and training set distribution verification.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental setup

4.1.1 dataset

This study employs two representative multimodal fake news datasets, Twitter and Weibo, for experimental validation. As shown in Table 1, the Twitter dataset [14] contains 13924 samples, including 7898 fake news items and 6026 authentic ones with 514 images. Its 514 images are shared visual assets, with multiple event or theme-related tweets referencing the same image. This aligns with real-world social media dissemination as a single visual asset is often reused across texts, reflecting the dataset's inherent design rather than a statistical discrepancy. In contrast, the Weibo dataset [15] includes 9528 samples, including 4749 fake news items and 4779 authentic ones, with one unique image per sample. Both datasets maintain a balanced distribution of fake and authentic information, providing a robust foundation for model training and evaluation.

Table 1. Statistical information on the dataset.

Method	Weibo	Twitter
# of Fake News	4749	7898
# of Real News	4779	6026
# of Images	9528	514

To ensure model robustness and prevent overfitting, this paper employs a rigorous data partitioning and validation strategy. Specifically, the two datasets are divided into training, validation, and test sets in a 7:1:2 ratio. The training set is used for model parameter learning, the validation set for hyperparameter tuning and early stopping, and the test set for final performance evaluation. To further validate model stability, a 5-fold cross-validation strategy is also adopted.

4.1.2 Data preprocessing

To ensure data quality and consistency, we designed a comprehensive data preprocessing workflow. As illustrated in Figure 5, raw data comprising textual content and accompanying images are initially collected from social media platforms, including Twitter and Weibo. Subsequently, text cleaning is performed to remove irrelevant symbols, URL links, and extraneous information, followed by text annotation. The images then undergo preprocessing operations, including size normalization and format conversion. Entity recognition algorithms are subsequently applied to extract key entities from both textual

and visual content. Finally, the feature extraction module generates standardized feature representations to facilitate model training.

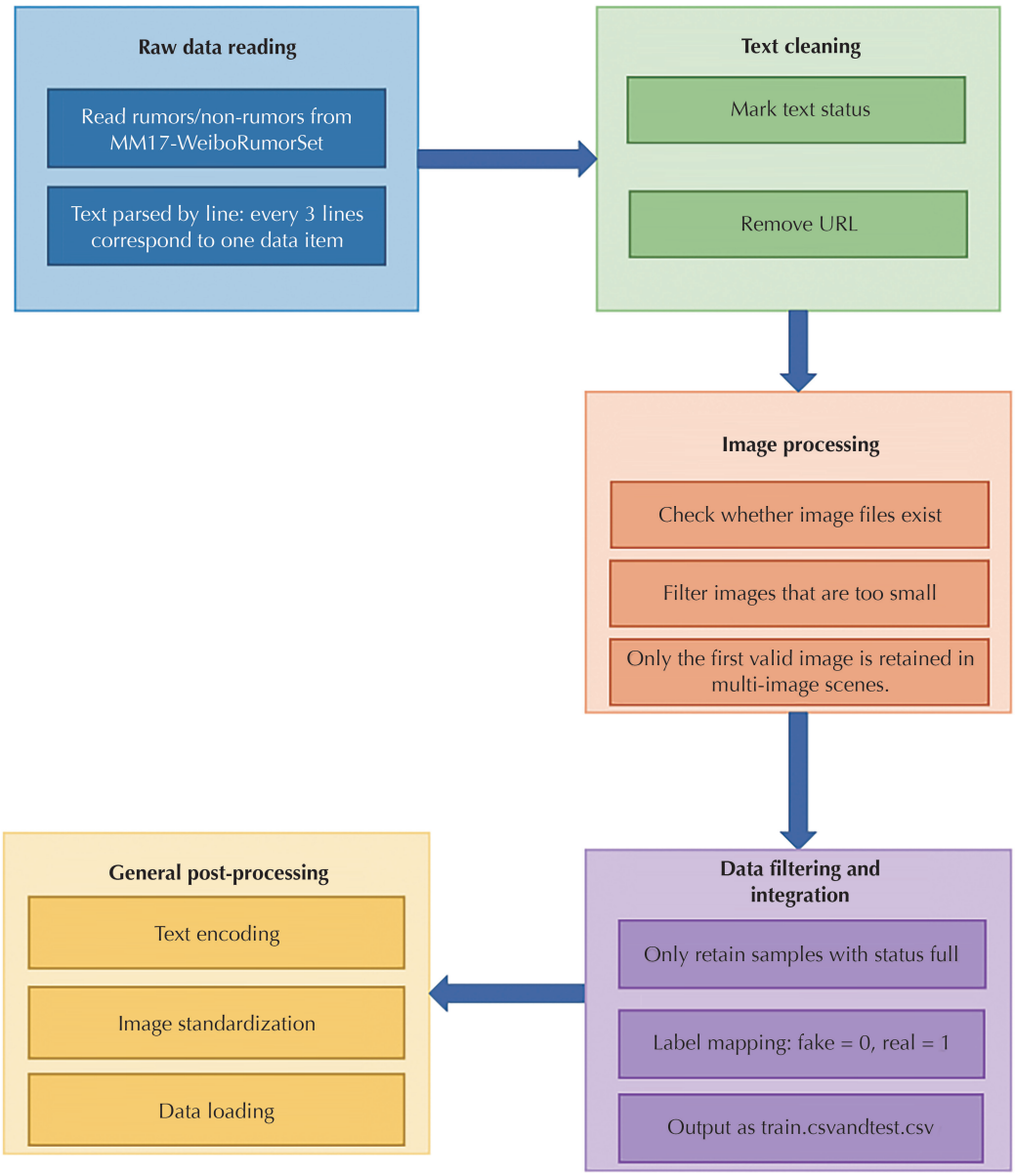


Figure 5. Data preprocessing process.

4.1.3 baseline model

To comprehensively evaluate the performance of EKAMF-Net, we selected several representative single-modal and multimodal fake news detection models as baselines:

(1) CNN uses a two-layer convolutional neural network architecture to learn feature representations for identifying fake news [43].

(2) VS focuses on learning different image distribution patterns between real and fake news by extracting various visual statistical features [44].

(3) att_RNN proposes a recurrent neural network based on an attention mechanism, integrating multimodal features such as text, visual, and social context. To ensure fairness in experimental comparisons, we removed social context features [45].

(4) EANN extracts event-invariant features through event classification tasks and uses these features to detect fake news in newly emerging events [19].

(5) MVAE uses a variational autoencoder to learn shared feature vectors from multimodal inputs and performs fake news detection based on these feature vectors [14].

(6) SAFE converts visual content into textual descriptions and detects fake news by exploring the similarity relationship between the textual content and the generated descriptions [22].

(7) HMCAN proposes a hierarchical multimodal contextual attention network that fuses the outputs of different layers of pre-trained BERT with visual features through a cross-attention mechanism [45].

(8) MCNN extracts text features, visual semantic features, and visual tampering features to explore the consistency of multimodal content [46].

(9) CAFÉ treats cross-modal ambiguity as a gating mechanism, adaptively aggregating unimodal features or capturing cross-modal correlations [5].

(10) CMMTN proposes a cross-modal feature alignment architecture based on a masked transformer, effectively filtering noise interference between text and visual components. This framework integrates a BERT-based text encoder with a ResNet visual representation module, employing a progressive positive-sample-unlabeled learning strategy to achieve fake news identification [47].

(11) CCGN employs graph neural network technology in the visual domain to eliminate background interference, narrows the semantic gap between modalities through a consistency comparison mechanism, and achieves feature fusion via an enhanced cross-modal attention mechanism [48].

4.1.4 Evaluation criteria

To comprehensively and objectively evaluate the performance of the model, the following evaluation indicators are used:

Accuracy (Acc): The proportion of correctly classified samples to the total number of samples, calculated using the following formula:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{28}$$

Precision (Prec): The proportion of samples that are fake news among samples predicted to be fake news. The calculation formula is:

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{29}$$

Recall (Rec): The proportion of samples that are correctly predicted as fake news among samples that are fake news. The calculation formula is:

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{30}$$

F1 score (F1-Score, F1): The harmonic mean of precision and recall, comprehensively reflecting the classification performance of the model. The calculation formula is:

$$\mathrm{F1} = \frac{2 \times \mathrm{Prec} \times \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}} \tag{31}$$

The confusion matrix parameters are defined as follows: TP (True Positive) represents samples that are actually fake news and correctly classified as fake news; TN (True Negative) represents samples that are authentic and correctly classified as authentic; FP (False Positive) represents samples that are authentic but incorrectly classified as fake news; and FN (False Negative) represents samples that are actually fake news but incorrectly classified as authentic.

4.2 Overall performance comparison test

To validate the overall performance of the EKAMF-Net model, it was compared with the above baseline models in experiments using Twitter and Weibo datasets. The experimental results are shown in Table 2.

Table 2. Performance comparison of each model on Twitter and Weibo datasets (F1 scores are class-specific for Fake News and Real News respectively. To comprehensively reflect the model's balanced detection capability for both classes, we adopt the macro-average F1 as the final evaluation metric, which is obtained by taking the arithmetic mean of the F1 scores of the two classes.).

Method	Weibo							Twitter						
	Acc	Fake News			Real News			Acc	Fake News			Real News		
		Pre	Rec	F1	Pre	Rec	F1		Pre	Rec	F1	Pre	Rec	F1
CNN (2017)	0.740	0.736	0.756	0.744	0.747	0.723	0.735	0.549	0.508	0.597	0.549	0.598	0.509	0.550
VS (2016)	0.726	0.732	0.712	0.722	0.720	0.740	0.730	0.617	0.635	0.644	0.639	0.639	0.630	0.634
att_RNN (2017)	0.772	0.854	0.656	0.742	0.720	0.889	0.795	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN (2018)	0.782	0.827	0.697	0.756	0.752	0.863	0.804	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE (2019)	0.824	0.854	0.769	0.809	0.802	0.875	0.837	0.745	0.801	0.719	0.758	0.689	0.777	0.730
SAFE (2020)	0.816	0.818	0.815	0.817	0.816	0.818	0.817	0.766	0.777	0.795	0.786	0.752	0.731	0.742
MCNN (2021)	0.823	0.858	0.801	0.828	0.787	0.848	0.816	0.784	0.778	0.781	0.779	0.790	0.787	0.788
CAFE (2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837	0.806	0.807	0.799	0.803	0.805	0.813	0.809
HMCAN (2021)	0.885	0.920	0.845	0.881	0.856	0.926	0.890	0,897	**0.971**	0.801	0.878	0.853	0.979	0.912
CMMTN (2024)	0.889	0.886	0.893	0.889	0.892	0.885	0.893	0.903	0.870	0.917	0.892	0.927	0.881	0.903
CCGN (2025)	0.908	0.922	0.894	0.908	0.894	0.922	0.908	0.906	0.961	0.748	0.841	0.886	**0.984**	0.933
EKAMF-Net	**0.924**	**0.928**	**0.910**	**0.918**	**0.907**	**0.928**	**0.917**	**0.930**	0.929	**0.921**	**0.925**	**0.937**	0.931	**0.934**

As shown in Table 2, multimodal models consistently outperform unimodal models on both datasets, indicating that integrating textual and visual features provides richer information for fake news detection. Among the multimodal models, EKAMF-Net demonstrates superior performance. On the Twitter dataset, EKAMF-Net achieved an macro-average F1 of 93.0%, representing a 3 percentage point improvement over the best baseline (CCGN). On the Weibo dataset, EKAMF-Net achieved an macro-average F1 of 91.8%, a 1 percentage point improvement over the baseline. These results demonstrate that EKAMF-Net effectively enhances multimodal fake news detection through entity knowledge enhancement and multi-scale adaptive fusion strategies.

Compared to other multimodal models, EKAMF-Net offers several key advantages. The model employs multi-scale feature extraction and dynamic threshold adaptive fusion, enabling superior capture of fine-grained textual and visual information while adjusting fusion weights based on cross-modal consistency strength. Additionally, EKAMF-Net incorporates an entity knowledge enhancement module through Wikipedia knowledge base integration, providing more targeted entity authenticity assessment and richer

factual information. The multi-scale adaptive fusion strategy addresses hierarchical differences in features and better handles complex textual-visual relationships, improving both detection performance and generalization capability across different cross-modal relationship states.

4.3 Ablation Experiments

To validate the effectiveness of each component in EKAMF-Net, we conducted ablation experiments. We systematically removed individual components: the entity knowledge enhancement module (-EK), multi-scale feature extraction module (-MS), dynamic threshold learning module (-DT), and adaptive fusion classification module (-AF), creating four variant models. Performance evaluation was conducted on both Twitter and Weibo datasets. Table 3 presents the experimental results, while Figure 6 illustrates the F1 score convergence of EKAMF-Net across training iterations for both datasets.

Table 3. Ablation experiment results (%).

Model	Dataset	Acc	Prec	Rec	F1
EKAMF-Net	Twitter	**93.0**	**93.3**	**91.1**	**93.0**
	Weibo	**92.4**	**92.4**	**91.5**	**91.8**
EKAMF-Net (-EK)	Twitter	90.1	89.5	88.9	89.2
	Weibo	87.0	86.3	85.6	85.9
EKAMF-Net (-MS)	Twitter	91.5	90.9	90.3	90.6
	Weibo	88.6	87.9	87.2	87.5
EKAMF-Net (-DT)	Twitter	92.1	91.5	90.8	91.1
	Weibo	89.8	89.1	88.4	88.7
EKAMF-Net (-AF)	Twitter	91.8	91.2	90.5	90.8
	Weibo	89.2	88.5	87.8	88.1

Table 3 demonstrates the effectiveness of each component through ablation analysis. Removing the entity knowledge enhancement module (-EK) results in F1 score decreases of 3.8 and 5.9 percentage points on Twitter and Weibo datasets, respectively, demonstrating its critical role in integrating entity factual information and enhancing fake news identification capabilities. Removing the multi-scale feature extraction module (-MS) decreases F1 scores by 2.4 and 4.3 percentage points, respectively, confirming its importance in capturing multi-granular textual and visual information for enhanced detection performance. Removing the dynamic threshold learning module (-DT) results in F1 score reductions of 1.9 and 3.1 percentage points, respectively, validating its role in learning adaptive thresholds based on cross-modal consistency for accurate feature fusion. Removing the adaptive fusion classification module (-AF) leads to F1 score decreases of 2.2 and 3.7 percentage points, respectively, highlighting its ability to adaptively adjust fusion strategies based on cross-modal relationships.

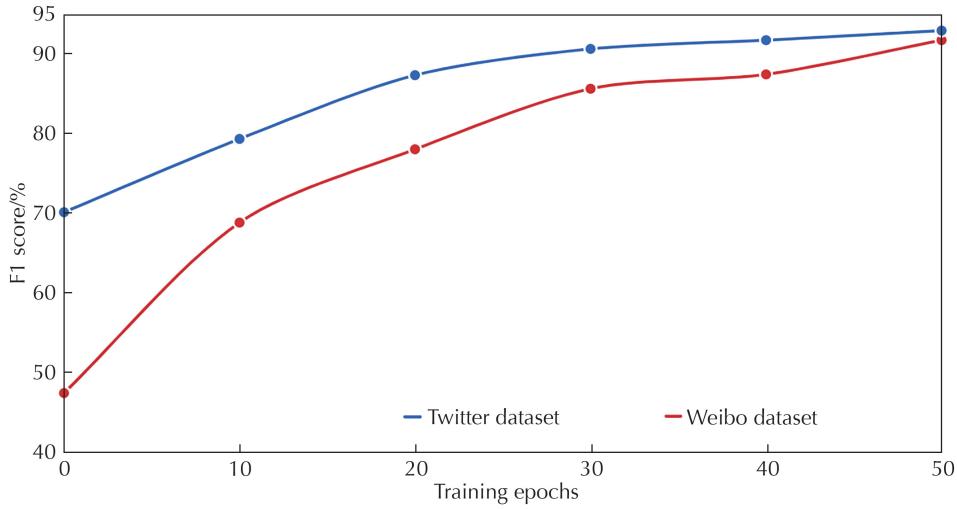


Figure 6. F1 score of EKAMF-Net on two datasets as a function of training iterations.

The ablation experiments demonstrate that all components contribute significantly to EKAMF-Net's performance, with the entity knowledge enhancement showing the largest impact (3.8–5.9 percentage points), followed by multi-scale feature extraction (2.4–4.3 percentage points) and adaptive fusion (2.2–3.7 percentage points). These components work synergistically to enhance multimodal fake news detection capabilities.

4.4 Performance Analysis of Different Types of Fake News Detection

To further analyze EKAMF-Net's performance, we categorized fake news into three types: semantic conflict (inconsistency between textual and visual semantics), entity falsification (fictitious entities in textual or visual content), and mixed deception (combining both semantic conflict and entity falsification). We evaluated the detection performance of each model across these fake news categories on both Twitter and Weibo datasets. Table 4 presents the results.

Table 4 demonstrates that EKAMF-Net achieves optimal performance across different fake news categories. EKAMF-Net shows particular strength in entity-based fake news detection, achieving F1 scores 4.7 percentage points higher than the best baseline (CAFÉ) across both datasets. This advantage stems from the entity knowledge enhancement module, which links textual and visual entities to Wikipedia to verify factual information and identify fictional entities. For semantic conflict fake news, EKAMF-Net demonstrates strong detection performance through its dynamic threshold adaptive fusion module, which accurately captures semantic consistency between textual and visual content. For mixed deception involving both semantic conflicts and entity falsification, detection becomes more challenging. However, EKAMF-Net still achieves robust results through the collaborative operation of its three modules.

Table 4. Detection performance of different models for different types of fake news F1 score, %.

Model	Dataset	Semantic conflict type	entity falsification type	Hybrid deception type
HMCAN	Twitter	**89.2**	85.3	87.6
	Weibo	**86.8**	82.9	84.7
MVAE	Twitter	**90.5**	86.7	88.9
	Weibo	**88.1**	84.2	86.0
MCNN	Twitter	**90.8**	87.2	89.3
	Weibo	**88.4**	84.7	86.4
CAFE	Twitter	**91.2**	87.6	89.7
	Weibo	**88.8**	85.1	86.8
EKAMF-Net	Twitter	93.5	92.3	**93.6**
	Weibo	**92.0**	89.8	91.0

4.5 Parameter sensitivity analysis

To investigate the impact of key parameters on model performance, a parameter sensitivity analysis was conducted. The attention weight coefficient α during entity knowledge feature injection and the weight coefficients of each loss term in the loss function were selected as the objects of study. The effects of different parameter values on the model's F1 score were tested on the Twitter dataset. Attention weight coefficient α. This parameter controls the extent to which entity knowledge features are injected into multi-scale features, with a value range of 0.1–0.9 and an interval of 0.1. The experimental results are shown in Figure 7.

As shown in Figure 7, the model's F1 score reaches its maximum when $\alpha = 0.5$. When α is too small, the entity knowledge features are insufficiently injected, failing to effectively enhance the factuality of multi-scale features. When α is too large, the interference of entity knowledge features on multi-scale features increases, affecting the model's ability to capture semantic information. Therefore, in subsequent experiments, $\alpha = 0.5$ is selected as the optimal parameter.

The weight coefficients of each loss term in the loss function are studied to investigate the impact of the weight coefficients of the classification loss ($\mathcal{L}_{cls}$), contrast loss ($\mathcal{L}_{cons}$), entity alignment loss ($\mathcal{L}_{ent}$), and threshold regularization loss ($\mathcal{L}_{thresh}$) on model performance. While keeping all other weight coefficients constant, we varied one weight coefficient within the range of 0–0.5, with an interval of 0.1, with the classification loss ($\mathcal{L}_{cls}$) set to 1 by default. The experimental results are shown in Figure 8.

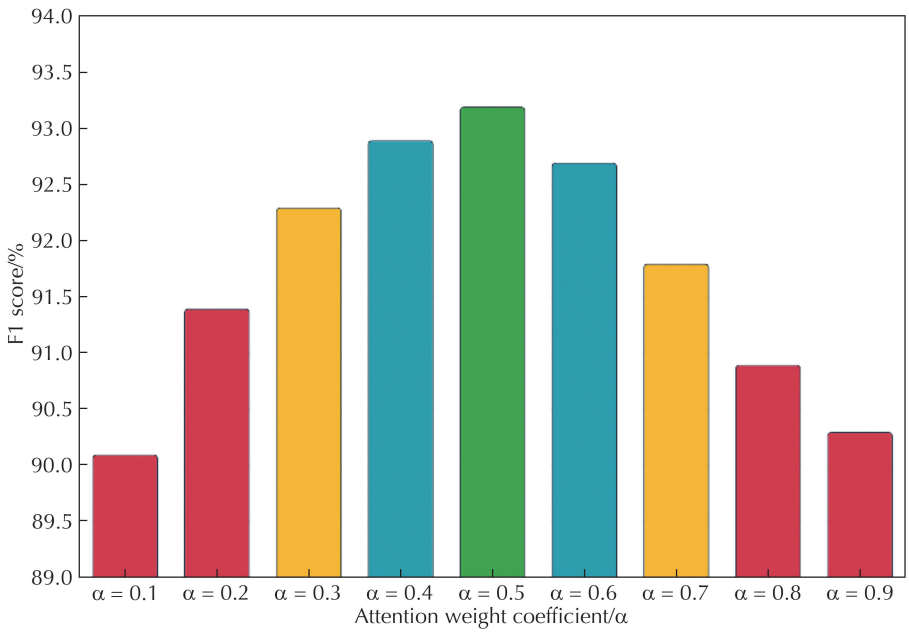


Figure 7. The effect of different attention weight coefficients α on the model's F1 score.

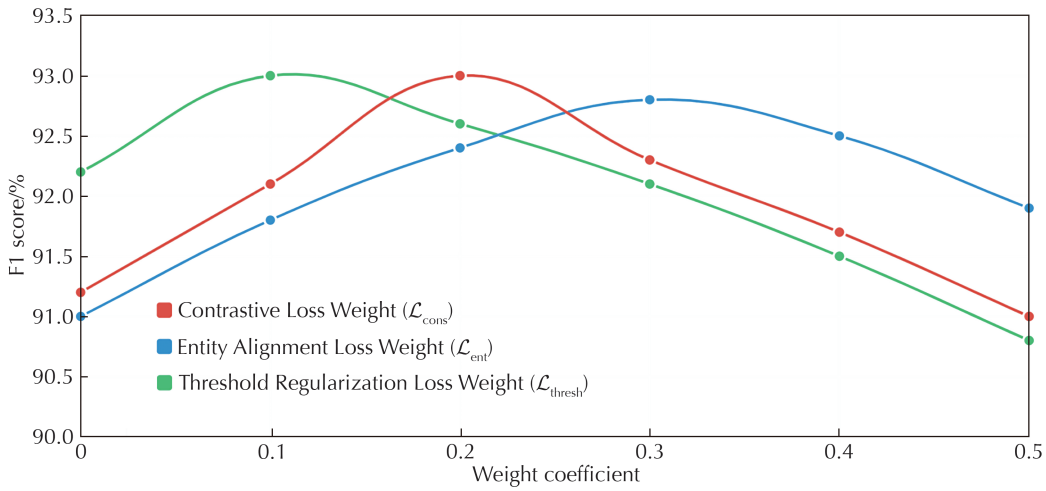


Figure 8. Impact of the weighting coefficients of each loss term in the loss function on the model's F1 score.

As shown in Figure 8, when the classification loss weight is 1.0 (default value), the contrast loss weight is 0.2, the entity alignment loss weight is 0.25, and the threshold regularization loss weight is 0.1, the model achieves the highest F1 score. This indicates that all loss terms play important roles in model training. Appropriately balancing the loss weights facilitates stable convergence and optimizes the final detection performance, as evidenced by the F1 score.

To more comprehensively validate the rationality of loss weight settings, we further conducted grid search experiments on multiple loss weight combinations. Specifically, we fixed the classification loss weight at 1.0 and the threshold regularization loss weight at 0.1. Under different threshold regularization loss weight configurations ($\mathcal{L}_{\text{thresh}}$ = 0.1, 0.2, 0.3, 0.4), systematically exploring the combination space of contrastive loss weight ($\mathcal{L}_{\text{cons}}$ = 0.1–0.3) and entity alignment loss weight ($\mathcal{L}_{\text{ent}}$ = 0.15–0.35). The experimental results are presented in the form of a heatmap in Figure 9.

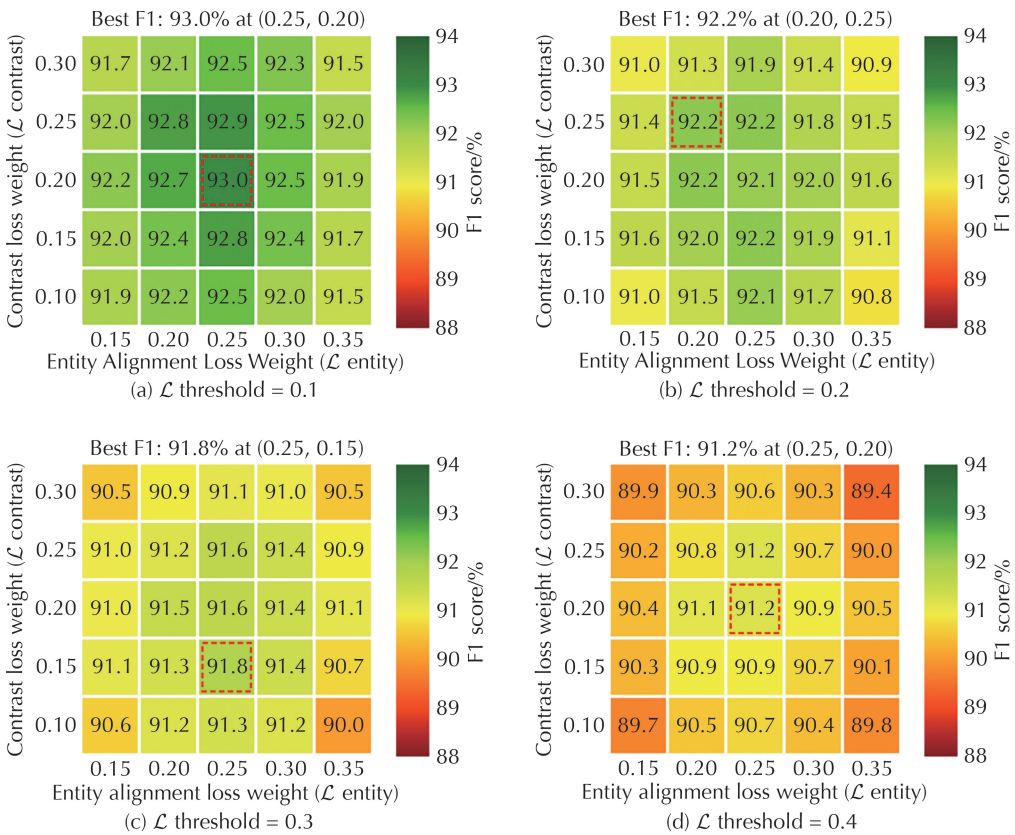


Figure 9. Heatmap of model F1 scores under different loss weight combinations. (The four subplots correspond to different threshold regularization loss weights (0.1, 0.2, 0.3, 0.4), with the optimal configuration region highlighted by the red dashed box.).

The F1 score heatmap of the model under different loss weight combinations in Figure 9 indicates that the weight configuration adopted in this paper (where $\mathcal{L}_{cls} = 1.0$, $\mathcal{L}_{cons} = 0.2$, $\mathcal{L}_{ent} = 0.25$, $\mathcal{L}_{thresh} = 0.1$) achieves an optimal global F1 score of 93%. Its rationality manifests in three key equilibrium points. First, the entity alignment loss weight of 0.25 resides in a optimal balance zone: below 0.2 leads to insufficient entity knowledge injection, while above 0.3 causes entity features to overshadow semantic features, disrupting multimodal feature balance. Second, the contrastive loss weight of 0.2 effectively serves as an auxiliary constraint without excessively limiting model representation capabilities, unlike values above 0.5. Finally, the threshold regularization weight of 0.1 preserves the flexibility of dynamic threshold learning while avoiding the overly restrictive effects of larger values like 0.4. This configuration achieves optimal synergy among entity knowledge enhancement, semantic feature extraction, and cross-modal fusion.

4.6 Analysis of Error Cases

To understand EKAMF-Net's limitations, we analyzed error cases from the test set. Error cases can be categorized into three main types: First, errors in identifying emerging entities. For recently emerged entities, Wikipedia's delayed updates prevent the entity knowledge enhancement module from obtaining effective features, reducing detection performance. For example, when analyzing news about a newly established organization lacking Wikipedia coverage, the model cannot perform entity verification, resulting in detection errors. Second, cross-language entity alignment errors. In Chinese datasets such as Weibo, entities often have names differing from their English Wikipedia counterparts, causing alignment errors during entity linking and reducing feature accuracy. For example, Chinese technology company names may correspond to multiple similar English Wikipedia entries, creating linking ambiguity and affecting detection accuracy. Finally, there are complex semantic understanding errors. For news content containing complex semantics such as metaphors and satire, the model has difficulty understanding their true meaning, leading to incorrect semantic consistency judgments. For example, satirical news may appear textually and visually consistent but contain hidden semantic conflicts that the model fails to identify, resulting in detection errors.

To quantitatively analyze the distribution of various error types and their impact on model performance, we systematically classified and statistically analyzed all error samples in the test set. Table 5 shows the proportion of each error type within the total error samples and their effect on overall performance.

As shown in Table 5, errors in identifying emerging entities are the primary factor affecting model performance, accounting for over 40% on average across both datasets and causing an approximately 3.8 percentage point drop in F1 scores. Cross-language entity alignment errors rank second, particularly prominent on the Weibo Chinese dataset, resulting in a 2.9% performance loss. Although errors in complex semantic understanding represent a relatively low proportion, they still account for 25.7% on the Twitter dataset.

Table 5. Error Type Distribution and Performance Impact Statistics Note: The performance reduction value indicates the percentage decrease in F1 score caused by this type of error.

Error Type	Twitter Dataset		Weibo Dataset		Average Performance Decrease (F1 Score %)
	Number of errors	Percentage (%)	Number of errors	Percentage (%)	
Emerging Entity Recognition Error	89	**42.4**	82	**41.8**	3.8
Cross-language entity alignment error	67	31.9	76	38.8	2.9
Complex Semantic Misunderstanding	54	25.7	38	19.4	1.6
Total	**210**	**100.0**	**196**	**100.0**	**8.3**

To further investigate the relationship between the proportion of emerging entities and detection accuracy, we divided the test set into different intervals based on the proportion of emerging entities and calculated the detection accuracy for each interval. The results are shown in Figure 10. As the proportion of emerging entities in the sample increases, the model's detection accuracy exhibits an approximately linear decline; When the proportion exceeded 15%, the rate of decline accelerated. Both datasets exhibited consistent downward trends, validating the universality of the emerging entity challenge. Based on this analysis, future improvements can be pursued through the following approaches. First, strengthen

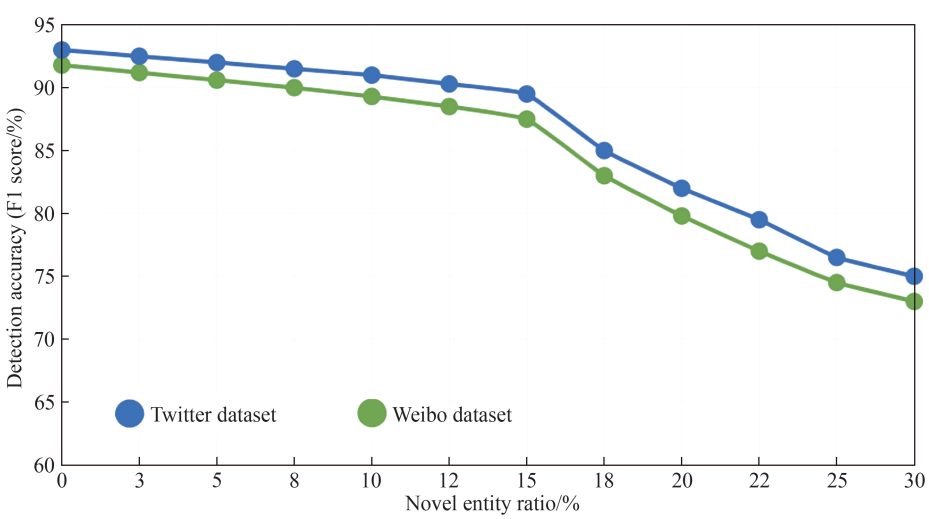


Figure 10. Relationship Between Proportion of Emerging Entities and Detection Accuracy.

real-time tracking of emerging entities and knowledge base updates to broaden entity coverage. Second, optimize cross-language entity alignment algorithms to improve entity linking precision. Third, incorporate more advanced natural language processing techniques to strengthen the model's ability to comprehend complex semantics.

5. CONCLUSION

This paper addresses challenges in multimodal fake news detection, including complex cross-modal semantic associations, insufficient entity fact verification capabilities, and poor dynamic adaptability of fusion strategies. It proposes a decoupled multimodal fake news detection model(EKAMF-Net) based on entity knowledge augmentation and multi-scale adaptive fusion. This model adopts a decoupled architecture, achieving precise detection of multimodal fake news through the collaborative operation of three core modules: multi-scale feature extraction, entity knowledge augmentation, and dynamic threshold adaptive fusion. The multi-scale feature extraction module designs distinct mechanisms for text and images, leveraging RoBERTa and CLIP-ViT-L/14 models to extract features from different levels, comprehensively capturing textual semantic information and visual information from images. The entity knowledge enhancement module associates entities in text and images with Wikipedia knowledge bases to obtain factual information. By injecting entity knowledge features into the multi-scale features, it enhances the model's ability to judge entity authenticity. The dynamic threshold adaptive fusion module dynamically adjusts fusion weights based on the consistency strength of cross-modal features, enabling effective integration of multi-scale features and improving the model's adaptability across different scenarios. Experimental results demonstrate that the EKAMF-Net model achieves F1 scores of 93.0% and 91.8% on the Twitter and Weibo datasets, respectively, representing significant improvements over other baseline models. Ablation experiments validated the effectiveness of each model component. Performance analysis across different fake news types revealed the model's distinct advantage in detecting entity-based fake news. Parameter sensitivity analysis determined optimal parameter settings, while error case analysis identified model limitations and future improvement directions.

In recent years, large multimodal foundational models such as GPT-4V and Gemini have demonstrated formidable capabilities across multiple tasks. This study opts to build task-specific architectures based on medium-scale pre-trained models, primarily driven by the following considerations. First, real-time detection scenarios impose stringent demands on inference efficiency. Detecting fake news on social media requires rapid responses within massive data streams, whereas large models typically exhibit inference latencies measured in seconds—far exceeding the millisecond-level processing requirements. EKAMF-Net employs a lightweight design enabling high-throughput inference on a single consumer-grade GPU, making it more suitable for practical deployment scenarios. Second, the decoupled architecture offers interpretability advantages. The end-to-end inference process of general-purpose large models is often a "black box," whereas this paper's decoupled design makes the model's decision-making process transparent and traceable. The entity knowledge module explicitly quantifies entity fact conflicts, the dynamic threshold module clearly delineates cross-modal consistency states, and the fusion module provides a basis for fine-grained weight allocation. This white-box characteristic is crucial for applications

requiring human review and result interpretation. Third, the universality and transferability of the modular mechanism. The three core mechanisms proposed in this paper—entity knowledge augmentation, multi-scale feature decoupling, and dynamic threshold fusion—are architecture-agnostic. Ablation experiments demonstrate that integrating these mechanisms into other baseline models also yields performance improvements, validating the universal value of the design approach. These mechanisms can also be transferred to large model frameworks in the future to construct hybrid architectures combining large model semantic understanding with specialized module-based structured reasoning. Fourth, cost sustainability for long-term operation. The invocation cost of commercial large model APIs increases linearly with sample size, posing a significant economic burden for continuously operating detection systems. Task-specific models exhibit extremely low marginal inference costs after training, offering greater economic viability. It is crucial to emphasize that this research does not negate the value of large models but advocates selecting the optimal solution for specific application scenarios. Large models excel in zero-shot generalization and complex reasoning, while task-specific models hold advantages in inference efficiency, interpretability, and cost control. These approaches complement each other: large models handle complex boundary cases, while lightweight models rapidly screen clear-cut instances, enabling a layered detection architecture. Future work will explore this hybrid strategy to achieve a better balance between performance and efficiency. In summary, the EKAMF-Net model offers a novel approach and methodology for multimodal fake news detection. By integrating entity knowledge enhancement with a multi-scale adaptive fusion strategy, the model achieves significant performance gains. Through continuous refinement and optimization, the model holds promise for further enhancing its capabilities and practicality, thereby making greater contributions to combating multimodal fake news and safeguarding societal information security.

AUTHOR CONTRIBUTIONS

Ailin Li and Wendi Zhu were responsible for conducting the experiments and writing the main content of the paper. Zhihang Wang undertook data collection and wrote part of the paper. Jizong Wang was in charge of drawing some charts and contributing to the paper. Wei Sheng handled data analysis and wrote part of the paper.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGEMENTS AND FUND

This work is supported by the Gansu University of Political Science and Law University-Level Key Research Project (GZF2025XZD20) and the National Natural Science Foundation of China (62366046). 2025 National College Students' Innovation Training Program Projects (202511406025).

REFERENCES

[1] S. Kumar and K. Carley, "Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, 2019, pp. 5047–5058, doi:10.18653/v1/p19-1498.

[2] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Understand.*, vol. 223, p. 103525, Oct. 2022, doi:10.1016/j.cviu.2022.103525.

[3] Y. Zhong, J. Shi, J. Yang, C. Xu, and Y. Li, "Learning to generate scene graph from natural language supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 1803–1814, doi:10.1109/ICCV48922.2021.00184.

[4] L. Peng, S. Jian, Z. Kan, L. Qiao, and D. Li, "Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection," *Inf. Process. Manage.*, vol. 61, no. 1, p. 103564, Jan. 2024, doi:10.1016/j.ipm.2023.103564.

[5] Y. Chen et al., "Cross-modal ambiguity learning for multimodal fake news detection," in *Proc. ACM Web Conf.*, Lyon, France, Apr. 2022, pp. 2897–2905, doi:10.1145/3485447.3511968.

[6] X. Fang et al., "NSEP: Early fake news detection via news semantic environment perception," *Inf. Process. Manage.*, vol. 61, no. 2, p. 103594, Mar. 2024, doi:10.1016/j.ipm.2023.103594.

[7] L. Han, X. Zhang, Z. Zhou, and Y. Liu, "A multifaceted reasoning network for explainable fake news detection," *Inf. Process. Manage.*, vol. 61, no. 6, p. 103822, Nov. 2024, doi:10.1016/j.ipm.2024.103822.

[8] P. Ferragina and U. Scaiella, "Fast and accurate annotation of short texts with Wikipedia pages," IEEE Softw., vol. 29, no. 1, pp. 70–75, Jan. 2012, doi:10.1109/MS.2011.122.

[9] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, no. 10, p. e20382, Oct. 2023, doi:10.1016/j.heliyon.2023.e20382.

[10] P. K. Verma, P. Agrawal, V. Madaan, and R. Prodan, "MCred: Multi-modal message credibility for fake news detection using BERT and CNN," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 8, pp. 10617–10629, Jul. 2022, doi:10.1007/s12652-022-04338-2.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol.*, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi:10.18653/v1/N19-1423.

[12] K. Nemkul, "Use of bidirectional encoder representations from transformers (BERT) and robustly optimized BERT pretraining approach (RoBERTa) for Nepali news classification," *Tribhuvan Univ. J.*, vol. 39, no. 1, pp. 124–137, Jun. 2024, doi:10.3126/tuj.v39i1.66679.

[13] D. Wu, Y. Yang, D. Qiu, and G. Li, "TimeMixer-ME: A dual-memory enhanced architecture for time series forecasting," *Front. Comput. Intell. Syst.*, vol. 12, no. 1, pp. 35–42, Apr. 2025, doi:10.54097/g0504662.

[14] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. WWW Conf.*, San Francisco, CA, USA, 2019, pp. 2915–2921, doi:10.1145/3308558.3313552.

[15] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in Proc. 25th ACM Int. Conf. Multimedia, Mountain View, CA, USA, Oct. 2017, pp. 795–816, doi:10.1145/3123266.3123454.

[16] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, Honolulu, HI, USA, Jul. 2019, pp. 5644–5651, doi:10.1609/aaai.v33i01.33015644.

[17] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Anchorage, AK, USA, Jul. 2019, pp. 395–405, doi:10.1145/3292500.3330935.

[18] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes," in Proc. 25th Int. Conf. World Wide Web, Montreal, QC, Canada, Apr. 2016, pp. 591–602, doi:10.1145/2872427.2883085.

[19] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, London, U.K., Jul. 2018, pp. 849–857, doi:10.1145/3219819.3219903.

[20] X. Liao, J. Yin, M. Chen, and Z. Qin, "Adaptive payload distribution in multiple images steganography based on image texture features," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 897–911, Mar.-Apr. 2022, doi:10.1109/TDSC.2020.3004708.

[21] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "FAMM: Facial muscle motions for detecting compressed deepfake videos over social networks," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 12, pp. 7236–7251, Dec. 2023, doi:10.1109/TCSVT.2023.3278310.

[22] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Adv. Knowl. Discov. Data Min.*, Houston, TX, USA, 2020, pp. 354–367, doi:10.1007/978-3-030-47436-2_27.

[23] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, Virtual, 2021, pp. 8748–8763, doi:10.48550/arXiv.2103.00020.

[24] E. Choi, J. Ahn, X. Piao, and J.-K. Kim, "CroMe: Multimodal Fake News Detection Using Cross-Modal Tri-Transformer and Metric Learning," IEEE Access, vol. 13, pp. 197124–197132, 2025, doi: 10.1109/access.2025.3633841.

[25] J. Li, Y. Bin, J. Zou, J. Wei, G. Wang, and Y. Yang, "Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection," in Proc. ACM Multimed. Asia, Singapore, Dec. 2023, pp. 1–7, doi:10.1145/3595916.3626397.

[26] H. Liu, W. Wang, and H. Li, "Interpretable multimodal misinformation detection with logic reasoning," in *Findings Assoc. Comput. Linguist.: ACL 2023*, Toronto, ON, Canada, 2023, pp. 9781–9796, doi:10.18653/v1/2023.findings-acl.620.

[27] S.-Y. Lin, Y.-C. Chen, Y.-H. Chang, S.-H. Lo, and K.-M. Chao, "Text-image multimodal fusion model for enhanced fake news detection," *Sci. Prog.*, vol. 107, no. 4, Oct. 2024, doi:10.1177/00368504241292685.

[28] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu: ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223 (2019).

[29] Y. Wang, X. Jiang, D. Cheng, D. Li, and C. Zhao, "Learning hierarchical prompts with structured linguistic knowledge for vision-language models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, Vancouver, BC, Canada, Mar. 2024, pp. 5749–5757, doi:10.1609/aaai.v38i6.28387.

[30] J. Wang, Z. Zhu, C. Liu, R. Li, and X. Wu, "LLM-enhanced multimodal detection of fake news," *PLOS ONE*, vol. 19, no. 10, p. e0312240, Oct. 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0312240. Accessed: Dec. 18, 2025.

[31] G. Ma, C. Hu, L. Ge, and H. Zhang, "DSMM: A dual stance-aware multi-task model for rumour veracity on social networks," *Inf. Process. Manage.*, vol. 61, no. 1, p. 103528, Jan. 2024, doi:10.1016/j.ipm.2023.103528.

[32] L. Zhang, Y. Shi, and M. Cui, "A bi-level multi-modal fake generative news detection approach: from the perspective of emotional manipulation purpose," *Humanit. Soc. Sci. Commun.*, vol. 12, no. 1, Jul. 2025, doi:10.1057/s41599025-05223-x.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.

[34] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 2022, pp. 1048–1058, doi:10.1109/CVPR52688.2022.01167.

[35] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal fake news detection: A survey of methods for combining visual and textual information," *Information*, vol. 13, no. 6, p. 284, Jun. 2022, doi:10.3390/info13060284.

[36] Y. Liu et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019).

[37] Z. Yuan, Z. Yang, H. Ning, and X. Tang, "Multiscale knowledge distillation with attention based fusion for robust human activity recognition," *Sci. Rep.*, vol. 14, no. 1, May 2024, doi:10.1038/s41598-024-63195-5.

[38] Y. Li, X. Liao, and X. Wu, "Screen-shooting resistant watermarking with grayscale deviation simulation," *IEEE Trans. Multimed.*, vol. 26, pp. 10908–10923, 2024, doi:10.1109/TMM.2024.3415415.

[39] L. Fu, X. Liao, J. Guo, L. Dong, and Z. Qin, "WaveRecovery: Screen-shooting watermarking based on wavelet and recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3603–3618, Apr. 2025, doi:10.1109/TCSVT.2024.3510355.

[40] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, 2019, pp. 4593–4601, doi:10.18653/v1/P19-1452.

[41] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proc. 57th Annu. Meet. Assoc. Comput. Linguist.*, Florence, Italy, 2019, pp. 3651–3657, doi:10.18653/v1/P19-1356.

[42] M. Raghu et al., "Do vision transformers see like convolutional neural networks?" *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/9825534692720b944f08901ae88f7a0-Paper.pdf. Accessed: Dec. 18, 2025.

[43] K. Ma et al., "DC-CNN: Dual-channel convolutional neural networks with attention-pooling for fake news detection," *Appl. Intell.*, vol. 53, no. 7, pp. 8354–8369, Aug. 2022, doi:10.1007/s10489-022-03910-9.

[44] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblog news verification," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 598–608, Mar. 2017, doi:10.1109/TMM.2016.2617078.

[45] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Virtual, Jul. 2021, pp. 153–162, doi:10.1145/3404835.3462871.

[46] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Inf. Process. Manage.*, vol. 58, no. 5, p. 102610, Sep. 2021, doi:10.1016/j.ipm.2021.102610.

[47] J. Wang, S. Qian, J. Hu, and R. Hong, "Positive unlabeled fake news detection via multi-modal masked transformer network," *IEEE Trans. Multimedia*, vol. 26, pp. 234–244, 2024, doi:10.1109/TMM.2023.3263552.

[48] S. Cui, K. Duan, W. Ma, and H. Shinnou, "CCGN: Consistency contrastive-learning graph network for multi-modal fake news detection," *Multimedia Syst.*, vol. 31, no. 2, Feb. 2025, doi:10.1007/s00530-025-01715-7.