Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues

Peng Qi1,2,3, Juan Cao1,2, Xirong Li4, Huan Liu5, Qiang Sheng1,2, Xiaoyue Mi1,2, Qin He6, Yongbiao Lv6, Chenyang Guo6, Yingchao Yu6
1Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
2University of Chinese Academy of Sciences 3Institute of Artificial Intelligence, Hebi, China
4Key Lab of DEKE, Renmin University of China, Beijing, China 5Zhengzhou University, Zhengzhou, China
6Hangzhou ZhongkeRuijian Technology Co., Ltd., Hangzhou, China
{qipeng,caojuan,shengqiang18z,mixiaoyue19s}@ict.ac.cn, xirong@ruc.edu.cn, liuhuan_2012@hotmail.com,
{heqin,lvyongbiao,guochenyang,yuyingchao}@ruijianai.com

ABSTRACT
Recently, fake news with text and images have achieved more effective diffusion than text-only fake news, raising a severe issue of multimodal fake news detection. Current studies on this issue have made significant contributions to developing multimodal models, but they are defective in modeling the multimodal content sufficiently. Most of them only preliminarily model the basic semantics of the images as a supplement to the text, which limits their performance on detection. In this paper, we find three valuable text-image correlations in multimodal fake news: entity inconsistency, mutual enhancement, and text complementation. To effectively capture these multimodal clues, we innovatively extract visual entities (such as celebrities and landmarks) to understand the news-related high-level semantics of images, and then model the multimodal entity inconsistency and mutual enhancement with the help of visual entities. Moreover, we extract the embedded text in images as the complementation of the original text. All things considered, we propose a novel entity-enhanced multimodal fusion framework, which simultaneously models three cross-modal correlations to detect diverse multimodal fake news. Extensive experiments demonstrate the superiority of our model compared to the state of the art.

CCS CONCEPTS
• Information systems → Multimedia information systems; Social networks.

KEYWORDS
fake news detection; multimodal fusion; visual entity; social media

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1 INTRODUCTION
The rising prevalence of fake news and its alarming real-world impacts have motivated both academia and industry to develop automatic methods to detect fake news (i.e., designing a classifier to judge a piece of given news as real or fake) [8, 11, 21, 31, 35]. Traditional approaches [4, 15, 18, 19] typically focus on the textual content, which is the main description form of news events. With the recent evolution of fake news from text-only posts to multimedia posts with images or videos [3], approaches based on multimodal content demonstrate promising detection performance [7, 9, 23, 26, 32]. This paper targets multimodal fake news detection, which is utilizing information of multiple modalities (here, text and images) to detect fake news.

Despite recent advancements in developing multimodal models to detect fake news, existing works model the multimodal content insufficiently. Most of them only preliminarily model the basic semantics of the images as the complement of the text, ignoring the characteristics of multimodal fake news. Specifically, some prior arts [23, 26, 27] obtain the multimodal representations by simply concatenating the textual features with visual features extracted from VGG19 [22] that is pre-trained on ImageNet [5].

To make up for this omission, we explore three valuable text-image correlations in multimodal fake news, which provide diverse multimodal clues. a) Text and images have inconsistent entities, which is a potential indicator for multimodal fake news. Wrongly reposting outdated images is a typical way to make up multimodal fake news [1, 2, 20]. However, it is difficult to find both semantically pertinent and non-manipulated images to support these non-factual stories in fake news, causing the inconsistency between text and images. For example, as shown in Figure 1(a), the text describes a piece of news about “Dallas Jones” while the attached image is the arrest scene of another person. b) Text and images enhance each other by spotting the important features. News text and images are related in high-level semantics, and the aligned parts usually reflect the key elements of news. In this kind of multimodal fake news, the text provides main clues for detection, while images help select the key clues in the text. As Figure 1(b) shows, the Nazi flag in the image corresponds to the important entity “Nazi” in the text, which is the key controversial point of this news post. c) The embedded text in images...


We will briefly review existing works on multimodal fake news detection (see Table 1) and explain our novelties accordingly.

The commonly used multimodal fusion framework for detection is to extract general visual features from pre-trained VGG19 [22] and then simply concatenate them with textual features. Based on this framework, Wang et al. [26] imported the event classification as an auxiliary task of fake news classification to guide the learning of event-invariant multimodal features for better generalizability. Then, Wang et al. [27] proposed a meta neural process approach to detect fake news on emergent events. Dhruv et al. [7] revised this framework into a multimodal variational autoencoder to learn a shared representation of multimodal contents for classification. Singhah et al. [23] first imported pre-trained language models (that is BERT, here) into this multimodal framework. Despite the advancements made by these works, they ignore the complex cross-modal correlations in fake news, which limits the effectiveness of multimodal content in detection.

Wrongly reposting irrelevant images is a typical way to make up multimodal fake news, and thus some works focus on measuring
the multimodal consistency for detection. Zhou et al. [34] used the image captioning model to translate the images into sentences and then computed the multimodal inconsistency by measuring the sentence similarity between the original text and the generated image captions. However, the translation performance is limited by the discrepancy between the training corpus of the image captioning model and the real-world news corpus, which further impairs the performance of cross-modal consistency measurement. Xue et al. [29] transformed the textual and visual features into a common feature space by weight sharing and then computed the cosine similarity of transformed multimodal features. Nevertheless, it is still hard to capture the multimodal inconsistency because of the semantic gap between textual and visual features.

On the other hand, some researchers proposed well-designed methods to model multimodal mutual enhancement. Jin et al. [9] proposed a neuron-level attention mechanism, and Zhang et al. [32] used the attention mechanism and multi-channel CNN to fuse multimodal information. These two works focus on the unidirectional enhancement of multimodal content, that is, highlighting the important image regions under textual guidance. Further, Song et al. [24] utilized the co-attention transformer to model the bidirectional enhancement between text and images. Wang et al. [28] extracted objects of the images and then use GCN to model the correlation between words and object labels. Similarly, Li et al. [12] extracted objects and then used the Capsule network to fuse the nouns and visual features of these objects. Nevertheless, these methods ignore the cross-modal enhancement on high-level semantics.

To sum up, there are two main drawbacks of existing works: 1) They do not consider these three cross-modal correlations simultaneously, and totally ignore the text complementation between the original text and the embedded text, and 2) model the cross-modal correlations based on the basic semantic features of the images, ignoring the news-related high-level visual semantics. To address these issues, we explicitly extract the visual entities and model the multimodal inconsistency and enhancement based on the multimodal entities. Moreover, we extract the embedded text in the images and model the text complementation. All things considered, we design a unified framework to fuse these multimodal clues for detection.

3 ENTITY-ENHANCED MULTIMODAL FAKE NEWS DETECTION

3.1 Model Overview

The goal of the proposed EM-FEND framework is to predict whether the given news is real or fake by utilizing its text $T$ and the attached image $I$.

As shown in Figure 2, EM-FEND includes three modules to fuse diverse multimodal clues for fake news detection: 1) Multimodal feature extraction, which extracts the textual and visual

Our model is applicable to news that contains multiple images, but for simplification we assume that only a single image is present in a piece of news.
Table 1: Comparison between EM-FEND and the state of the art for multimodal fake news detection. These compared methods do not consider three cross-modal correlations at the same time.

| Methods      | Backbone       | Fusion       | Cross-modal Correlations |
|--------------|----------------|--------------|--------------------------|
|              | Text Image     |              | inconsistency enhancement text complementation |
| EANN[26]     | Text-CNN       | VGG19 concat | -                        |
| metaFEND[27] | Text-CNN       | VGG19 concat | -                        |
| MVAE[7]      | Bi-LSTM        | VGG19 varia- | -                        |
| SpotFake[23] | BERT           | VGG19 concat | -                        |
| SAFE[34]     | Text-CNN       | image2sentence concat multi-loss text-imagecaption | - |
| MCNN[29]     | BERT +Bi-GRU   | ResNet50 attention multi-loss text-visfea | - |
| attRNN[9]    | Bi-LSTM        | VGG19 neuron-level attention | - text->visfea |
| MKEMN[32]    | Bi-GRU         | VGG19 attention +multi-channel CNN +attention transformer +multi-channel CNN | - text<>visfea |
| CARMN[24]    | BERT           | VGG19 co-attention transformer | - text<>visfea |
| KMGCN[28]    | -              | YOLOv3 GCN | - text<>objects |
| EMAF[12]     | BERT           | Faster-RCNN Capsule | - text<>object fea |

EM-FEND(ours) BERT VGG19 +entity detector +OCR model co-attention transformer text-visentity text<>visfea text<>visentity +

entities, the embedded text in the image, and the visual CNN features (Section 3.2); 2) Multimodal feature fusion, which models three types of cross-modal correlations, including entity inconsistency, mutual enhancement, and complementation (Section 3.3); and 3) Classification, which uses the obtained multimodal representation to perform binary classification (Section 3.4). We will introduce the above modules in detail.

3.2 Multimodal Feature Extraction

3.2.1 Text Input.

Textual Entities. As a special narrative style, news usually contains named entities such as persons and locations. These entities are of importance in understanding the news semantics and also helpful in detecting fake news. Thus, we explicitly extract the person entities \( P_T \) and location entities \( L_T \) by recognizing corresponding proper nouns in the text. For better understanding the news events, we employ the part-of-speech tagging to extract all nouns as a general textual context \( C_T \).

Visual CNN Features. Following previous works, we adopt VGG19 to extract the visual features. Unlike these works, we fine-tune the pre-trained VGG19 on the given dataset to flexibly capture the low-level characteristics of the images from the specific data source to help detection. For example, the image quality is a powerful feature for distinguishing fake news and real news posts on social media, while it is less effective for detecting fake news articles on formal news sites. Then, we extract the visual features of the input image from the output of the last layer of VGG19. Considering that different regions in the image may show different patterns, we split the original image into \( 7 \times 7 \) regions, and then obtain the corresponding visual features sequence \( H_V = [r_1, ..., r_n], n = 49 \), where \( r_i \) represents the feature of the \( i \)-th region in the image.

Visual Entities. Similar to the text, news images also contain newsworthy visual entities, which are important for semantic understanding and fake news detection. Specifically, we extract four types of visual entities: 1) celebrities and landmarks; 2) organizations, such as Nazi, Buddhism and police, by detecting flags or clothes; 3) eye-striking visual concepts, such as violence, bloodiness, and disaster [20]; and 4) general objects and scenes. Due to the high accuracy requirements for pre-trained models and the lacking of relevant publicly available datasets, we use public APIs\(^3\) to detect visual entities instead of re-implement these models. Finally, we obtain the person entities \( P_V \), location entities \( L_V \), and other news-related visual concepts with corresponding probability as a more general image context \( C_V \).

Embedded Text. In addition to the original input text, text embedded in images is also important because it usually contains important information missed by the original text. We extract the embedded text \( O \) of the input image by applying the optical character recognition (OCR) model\(^4\).

3.3 Multimodal Feature Fusion

3.3.1 Text Complementation. As the main body of multimodal news, the text provides rich clues for the judgment of news credibility. For fake news in social media, in addition to the original text, the embedded text in images is also important in understanding the news semantics and providing clues for detection. In many situations, the key clues for detection lie in the embedded text, while the original text is just a comment about the news event. Therefore, the

\(^3\)https://ai.baidu.com/tech/imagecensoring, https://ai.baidu.com/tech/imagerecognition
\(^4\)https://ai.baidu.com/tech/ocr/general
Specifically, we feed the textual features \( H \) of multimodal feature heterogeneity. The aligned words and visual entities usually reflect the key elements of the news, and thus we feed the original textual features and visual entities’ embeddings could be fused in trained BERT to obtain their embeddings \( \mathbf{V}, \mathbf{E} \) and text-enhanced visual entities representation \( H_{VE-T} \).

**MCT between Textual Features and Visual CNN Features.** Visual entities focus on the local high-level semantics of the images, while ignoring the global low-level visual features. As a supplement, we use the multimodal co-attention transformer to model the correlations between textual features and visual CNN features. Specifically, we feed \( H_{T-V} \) and the visual CNN features \( H_V \) into the second co-attention transformer, obtaining the textual representation enhanced by both visual entities and visual CNN features \( H_{T-V(VE,V)} \) and text-enhanced visual representation \( H_{V-E-T} \).

We apply the average operation on the above features to obtain the final representation of the text and image, that is \( x_T \) and \( x_V \), respectively.

### 3.3.3 Entity Inconsistency Measurement

Multimodal entity inconsistency is a potential indicator for multimodal fake news. For example, if the person mentioned in the text is inconsistent with the recognized celebrity in the image, this news post may be fake with misused images (see Figure 1(a)). Motivated by Müller-Budack et al. [16], we measure the multimodal entity inconsistency of person, location, and a more general event context. There are two challenges for this measurement: the first one is the heterogeneity of textual and visual features. Unlike previous works that calculate the multimodal similarity in transformed [29] or visual feature spaces [16], we calculate the similarity of multimodal entities in textual feature space based on their word embeddings. Second, news text usually contains more entities and information than the accompanying images, and thus some textual entities could be without the aligned visual entities. Considering that fake news commonly tampers only one entity type to maintain credibility, we consider the multimodal news as entity inconsistent only when there are no aligned multimodal entities.

Taking person entity as an example, we define the cross-modal person similarity as the maximum similarity among all pairs of textual and visual person entities. Since neural networks have inevitable errors when detecting visual entities, the confidence is considered when computing the similarity. Formally, we define \( t \) and \( v \) as the feature vectors of the textual and visual entities. For a news post with \( T_P \) and \( V_P \), we calculate the cross-modal person similarity as

\[
  x_t^p = \max_{t \in T_P} \left( \sum_{v \in V_P} \rho(v) \frac{t \cdot v}{||t|| ||v||} \right),
\]

where \( \rho(v) \) represents the probability of visual entity \( v \). For news that lacks textual or visual entities, we set the similarity as 1 to indicate no effective clue about multimodal inconsistency for fake news detection. Similarly, we compute the cross-modal location similarity \( x_t^l \) and context similarity \( x_t^c \), and then concatenate them to form the entity consistency feature \( x_t = [x_t^p, x_t^l, x_t^c] \).

Finally, we concatenate the final representation of the text \( x_T \), that of visual entities \( x_{VE} \), that of the image \( x_V \), and the multimodal entity consistency feature \( x_T \) to obtain the final multimodal representation as

\[
  x_m = \text{concat}(x_T, x_{VE}, x_V, x_T).
\]
3.4 Classification

Till now, we have obtained the final multimodal representation \( x_m \), which models the input multimodal news from multiple perspectives. We use a fully connected layer with softmax activation to project the multimodal feature vector \( x_m \) into the target space of two classes: real and fake news, and gain the probability distributions:

\[
p = \text{softmax}(W x_m + b),
\]

where \( p = [p_0, p_1] \) is the predicted probability vector with \( p_0 \) and \( p_1 \) indicating the predicted probability of label being 0 (real news) and 1 (fake news), respectively. \( W \) is the weight matrix and \( b \) is the bias term. Thus, for each news post, the goal is to minimize the binary cross-entropy loss function as follows,

\[
L_p = -[y \log p_0 + (1 - y) \log p_1],
\]

where \( y \in \{0, 1\} \) denotes the ground-truth label.

4 EXPERIMENTS

In this section, we conduct experiments to evaluate the effectiveness of the proposed EM-FEND. Specifically, we aim to answer the following evaluation questions:

- **EQ1**: Can EM-FEND improve the classification performance of distinguishing multimodal fake and real news?
- **EQ2**: How effective are various visual features (especially the visual entities) and cross-modal correlations in improving the performance of EM-FEND?
- **EQ3**: How does EM-FEND perform in online fake news detection?

4.1 Datasets

To prove the generalization of the proposed EM-FEND, we conduct experiments on two real-world datasets of different languages.

4.1.1 Chinese Dataset. The Chinese dataset is constructed on the Chinese Sina Weibo microblogging platform by Jin et al. [9] and has been broadly used in existing works [7, 23, 26]. The fake news posts are verified by the official rumor debunking website of Weibo\(^5\), which serves as a reputable source to collect fake news posts in literature. The real news posts are collected from Weibo during the same period as the fake news and are verified by Xinhua News Agency, an authoritative news agency in China. This dataset has been preprocessed to ensure that each post corresponds to an image. In total, this dataset includes 4,749 fake news posts and 4,779 real news posts with corresponding images.

4.1.2 English Dataset. The English dataset is proposed by Yang et al. [30]. The fake news is crawled from news websites that are manually assessed as low credibility\(^6\). And the real news is crawled from well-known authoritative news websites such as the New York Times. After removing text-only news, non-English news, and news with unavailable images, we obtain 2,844 fake news articles and 2,825 real news articles, each corresponding to an image.

To prevent the model from overfitting on event topics, we first use the K-means algorithm to find the common events and split the data into training, validation and testing sets based on these event clusters to ensure that there is no event overlap among these sets [26]. The training, validation, and testing sets contain data approximately with a ratio of 3:1:1. We use the Accuracy (Acc.) and Precision (Prec.), Recall and F1 score of the fake-news class as evaluation metrics.

4.2 Implementation Details

We use the pre-trained BERT models\(^7\) (i.e., bert-base-chinese and bert-base-uncased) to obtain the textual representation. The maximum sequence length is 256 for both datasets. For models that are not based on BERT, we use publicly available Word2Vec models\(^8\) to obtain the word embeddings. For detecting textual entities, we use public API\(^9\) and the open-sourced library Spacy\(^10\) for Chinese and English news, respectively. In the co-attention transformer block, we employ 8 heads and the hidden size is set as 256 and 128 for EM-FEND and EM-FEND-base, respectively. The hidden size of LSTM in the EM-FEND-base is 128. We use a batch size of 64 instances in the training process. The model is trained for 100 epochs with early stopping to prevent overfitting. We use ReLU as the non-linear activation function and use Adam\(^10\) algorithm to optimize the loss function. The dropout rate is set as 0.3.

4.3 Comparison Methods

To validate the effectiveness of the proposed EM-FEND framework, we compare it with several representative methods including single-modality and multimodal methods as follows:

**Single-modality Methods**

- **Bi-LSTM**: uses a network based on the bidirectional LSTM to classify the given piece of news.
- **BERT**: uses a pre-trained BERT to obtain the representation of the given piece of news and a fully connected layer to make classifications.
- **VGG19**: fine-tunes VGG19 to model news images for classifications.

**Multimodal Methods**

- **attRNN** [9]: proposes an innovative RNN with an attention mechanism for effectively fusing multimodal features. In detail, it produces the joint features of text and social context by an LSTM network and fuses them with visual features by utilizing the neural-level attention from the outputs of the LSTM. For a fair comparison, we remove the part dealing with social context features.
- **MVAE** [7]: utilizes a multimodal variational autoencoder trained jointly with a fake news detector to learn a shared representation of multimodal content for fake news detection. It is composed of textual and visual encoders and corresponding decoders, and a fake news detector.
- **MKN** [32]: retrieves concepts of textual entities from external knowledge graphs and proposes a multi-channel word-knowledge-visual-aligned CNN for fusing multimodal information. The original model MKEMN uses an event memory

---

\(^5\)https://service.account.weibo.com
\(^6\)https://ai.baidu.com/tech/nlp_basic/lexical
\(^7\)https://github.com/google-research/bert
\(^8\)https://ai.tencent.com/ailab/nlp/en/embedding.html
\(^9\)https://github.com/mmmihaltz/word2vec-GoogleNews-vectors
\(^10\)https://github.com/mmihaltz/word2vec-GoogleNews-vectors

1https://www.kaggle.com/mrisdal/fake-news
We compare EM-FEND with representative methods introduced in Section 4.3. The results are presented in Table 2, from which we can draw the following observations:

**Table 2: Performance comparison for multimodal fake news detection on two real-world datasets.**

| Methods          | Acc. | Prec. | Recall | F1   |
|------------------|------|-------|--------|------|
| Bi-LSTM          | 0.785| 0.851 | 0.692  | 0.763|
| BERT             | 0.830| 0.977 | 0.675  | 0.798|
| VGG19            | 0.730| 0.789 | 0.626  | 0.698|
| attRNN[9]        | 0.808| 0.882 | 0.711  | 0.787|
| MVAE[7]          | 0.797| 0.827 | 0.751  | 0.787|
| MKNN[32]         | 0.805| 0.865 | 0.722  | 0.787|
| SAFE[34]         | 0.790| 0.886 | 0.665  | 0.760|
| EM-FEND-base (Ours) | 0.852| 0.841 | 0.853  | 0.847|
| SpotFake[23]     | 0.852| 0.854 | 0.850  | 0.852|
| CARMN [24]       | 0.865| 0.933 | 0.774  | 0.846|
| EM-FEND (Ours)   | **0.904**| 0.897 | **0.904**| **0.901**|
| Bi-LSTM          | 0.864| 0.877 | 0.843  | 0.859|
| BERT             | 0.873| 0.869 | 0.875  | 0.872|
| VGG19            | 0.773| 0.783 | 0.747  | 0.764|
| attRNN[9]        | 0.872| 0.861 | 0.882  | 0.871|
| MVAE[7]          | 0.879| 0.902 | 0.848  | 0.874|
| MKNN[32]         | 0.889| 0.846 | 0.929  | 0.886|
| SAFE[34]         | 0.909| 0.922 | 0.890  | 0.906|
| EM-FEND-base (Ours) | 0.943| 0.926 | 0.961  | 0.943|
| SpotFake[23]     | 0.899| 0.879 | 0.923  | 0.901|
| CARMN [24]       | 0.937| 0.934 | 0.940  | 0.937|
| EM-FEND (Ours)   | **0.975**| **0.978**| **0.973**| **0.975**|

**Table 3: Ablation study on various visual features.**

| Methods          | Methods          |
|------------------|------------------|
| Chinese          | English          |
|                  | Methods          |
|                  | Acc. | Prec. | Recall | F1   |
| EM-FEND          | **0.904**| 0.897 | **0.904**| **0.901**|
| w/o visual entities | 0.886 | 0.930 | 0.823  | 0.873|
| w/o OCR text     | 0.882 | 0.902 | 0.845  | 0.873|
| w/o FT VGG feature | 0.773 | 0.783 | 0.747  | 0.764|
| Em-FEND          | **0.975**| **0.978**| **0.973**| **0.975**|
| w/o visual entities | 0.953 | 0.954 | 0.950  | 0.952|
| w/o OCR text     | 0.970 | 0.967 | 0.972  | 0.969|
| w/o FT VGG feature | 0.970 | 0.954 | 0.988  | 0.971|

- **EM-FEND is much better than other methods on both datasets, no matter whether or not to adopt BERT as the textual feature extractor.** It validates that EM-FEND can effectively capture important multimodal clues that existing works ignore to detect fake news. Specifically, EM-FEND and EM-FEND-base outperform the corresponding state-of-the-art methods by at least 3.8 and 3.4 percentage points in accuracy, respectively.
- **Methods based on textual modality are better than the visual modality, proving that the text provides more rich clues than images.** Multimodal methods are generally better than methods based on single-modality, indicating the complementarity of multimodal features.
- **Pre-trained language models (i.e., BERT) can improve the performance of our method.** It is mainly due to the strong ability of transformers in modeling context and the abundant knowledge injected in the pre-trained models.

### 4.4 Performance Comparison (EQ1)

We design two groups of ablation experiments to evaluate the effectiveness of different components in EM-FEND. Specifically, we design several internal models for comparison, which are simplified variations of EM-FEND with certain visual features removed:

- **SAFE** [34]: translates the input image into a sentence, and computes the multimodal relevance based on the sentence similarity as the auxiliary loss for the fake news classification.
- **SpotFake** [23]: concatenates the textual and visual features obtained from pre-trained BERT and VGG19 respectively for classification.
- **CARMN** [24]: proposes a cross-modal attention residual network to fuse multimodal features. We use the pre-trained BERT to obtain the textual representation.

Considering that using pre-trained language models to extract textual features usually improves the detection performance of models even without significant changes on the model structure [6], we design a reduced variant of the proposed EM-FEND model to ensure the fairness of comparisons.

- **EM-FEND-base**: uses a Bi-LSTM with pre-trained Word2Vec models to replace BERT in EM-FEND when obtaining the textual features. The embeddings of textual and visual entities are also obtained by pre-trained Word2Vec models.

### 4.5 Ablation Study (EQ2)

We design two groups of ablation experiments to evaluate the effectiveness of different components in EM-FEND. Specifically, we design several internal models for comparison, which are simplified variations of EM-FEND with certain visual features removed:

- **w/o visual entities**: EM-FEND without the visual entities extraction, and the following co-attention transformer between textual features and visual entities and entity inconsistency measurement module.
- **w/o OCR text**: EM-FEND without the OCR text.
- **w/o fine-tuned (FT) VGG feature**: We extract visual features from pre-trained VGG19 without fine-tuning.

Similarly, we design the following variants of EM-FEND to prove the effectiveness of different cross-modal correlations:

- **w/o co-attention-ve**: EM-FEND without the co-attention transformer between textual features and visual entities.
- **w/o co-attention-vf**: EM-FEND without the co-attention transformer between textual and visual CNN features.
- **w/o entity inconsistency measurement**: EM-FEND without the entity inconsistency measurement module.

The results of the ablation study are reported in Table 3 and Table 4. We have the following observations:

1) **Visual Features**: All of these three visual features are important for fake news detection. However, the most important visual features on these two datasets are different: fine-tuned VGG features in the Chinese dataset and visual entities in the English dataset.
This phenomenon results from the differences in data sources between these two datasets. The Chinese dataset is collected from the social media platform, and thus the multimodal fake news is more likely to show characteristics of low image quality brought by wide propagation. Differently, the English dataset originates from the formal news websites, of which the news has high-quality and informative images. Thus, the high-level visual semantic features are more important than low-level visual features for detecting this kind of multimodal fake news. This phenomenon also proves the generalization ability of EM-FEND in detecting different types of multimodal fake news.

2) Cross-modal Correlations: Other than visual features, the various cross-modal correlations are also important for achieving the best performance of EM-FEND. If we remove one of them, the performance will drop by a certain degree. Specifically, i) the accuracy is lower than the complete model by at least 1.4 percentage points in accuracy when we replace the single co-attention transformer module with the average operation, proving that the co-attention transformer can effectively fuse multimodal features by capturing the multimodal alignment; ii) The influence of entity consistency is smaller than other cross-modal correlations, probably due to the sparsity of visual entities and the noises brought by entity detectors.

4.6 Robustness to Imbalanced Online Data (EQ9)

In real-world scenarios, the number of fake news is much lower than real news, which means that the online news data that needs to be detected is imbalanced. We collect news from an online fake news detection system like [33] during 9 months. After removing news posts without text or images and duplicated posts, we obtain 217 multimodal fake news posts and 3353 real news posts annotated by experts, with a ratio of 1:15 approximately. It is worth noting that it’s more difficult to distinguish these fake and real news than distinguishing that in datasets used in Section 4.1, because these real news posts originate from suspicious news and usually show typical patterns of fake news.

To evaluate the robustness of EM-FEND to imbalanced online data, we compare EM-FEND with CARMN [24], the best competitor to EM-FEND (Table 2), in the imbalanced dataset. Figure 4 shows the ROC curves of these two models, from which we observe that EM-FEND outperforms CARMN in online data.

### Table 4: Ablation study on various cross-modal correlations.

| Methods                      | Acc. | Prec. | Recall | F1    |
|------------------------------|------|-------|--------|-------|
| **Chinese**                  |      |       |        |       |
| EM-FEND                      | 0.904| 0.897 | 0.904  | 0.901 |
| w/o entity consistency       | 0.899| 0.932 | 0.849  | 0.889 |
| w/o co-attention-ve         | 0.890| 0.914 | 0.851  | 0.881 |
| w/o co-attention-vf         | 0.886| 0.901 | 0.855  | 0.878 |
| **English**                 |      |       |        |       |
| EM-FEND                      | 0.975| 0.978 | 0.973  | 0.975 |
| w/o entity consistency       | 0.962| 0.977 | 0.945  | 0.961 |
| w/o co-attention-ve         | 0.959| 0.953 | 0.966  | 0.959 |
| w/o co-attention-vf         | 0.930| 0.937 | 0.920  | 0.928 |

4.7 Case Study

In this part, we show some cases to intuitively show the behaviors of entity inconsistency measurement module in EM-FEND. Specifically, we list several representative multimodal fake news that are measured as low person consistency in Figure 5. It shows that this module can effectively measure the multimodal entity inconsistency as easily understanding explanations for the model’s decisions about fake news.

![Image of fake news examples](image_url)

**Figure 5:** Some fake news with low multimodal person consistency. In these cases, the person entity mentioned in the text is inconsistent with that recognized in the image.

5 CONCLUSION

In this paper, we find three valuable cross-modal correlations in multimodal fake news on social media, that is entity inconsistency, mutual enhancement and text complementation, which provides diverse multimodal clues. Also, we reveal the importance of visual entities in understanding news-related visual semantics and capturing these multimodal clues. Accordingly, we propose a novel entity-enhanced multimodal fusion framework named EM-FEND to simultaneously model three cross-modal correlations. Extensive experiments have proved the effectiveness of EM-FEND.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2017YFC08020604), and the National Natural Science Foundation of China (U1703261, 62172420).
