MKGRP-FND: False News Detection Based on Multimodal Knowledge Graph Relation Prediction

Anonymous ACL submission

Abstract

The proliferation of false news on social media poses a significant threat to public discussions and social stability. The existing knowledge-enhancement methods are mostly based on single-modal knowledge graphs and utilize knowledge to enhance the model's ability to understand images and capture semantic relationships. However, those methods require a large amount of feature computation costs and is difficult to avoid the loss of knowledge during the embedding and transformation process, which significantly affects the model's ability to detect details. To address these shortcomings, we firstly propose a novel false news detection framework based on multi-modal knowledge graph relation prediction(MKGRP-FND). This framework converts news content into a graph structure representation and transforms the multimodal false news detection task into a relationship prediction task in the knowledge graph, thereby enabling fine-grained authenticity checks at the entity and relationship levels without relying on the overall feature representation. Experiments on the PHEME public dataset show that MKGRP-FND can effectively identify cross-modal semantic inconsistency and factual anomalies, providing a reliable and understandable method for false news detection with good interpretability.

1 Introduction

A large amount of false multimodal content is widely spread on social platforms, posing a significant threat to public opinion and social stability.Multimodal False News Detection(MFND), which aims to ascertain the veracity of news content comprising both text and images, is a critical task for maintaining the integrity of digital ecosystems (Hu et al., 2023).

Early research mainly focuses on single-modal analysis using language features (Kaliyar et al., 2021), communication patterns (Yue et al., 2023), or user credibility (Adaji and Okpanachi, 2024). With the rise of multimedia, multimodal fusion models jointly model text and image features for classification (Wang et al., 2018; Qi et al., 2021). Although effective for detecting stylistic or emotional alterations, they lack external factual basis. Thus, for fact-based false statements that do not exhibit obvious abnormal low-level features, these models cannot provide clear, evidence-based conclusions and are prone to being influenced by contextually plausible deception (Xie et al., 2024; Li et al., 2025a).

(a) (b)

Figure 1: Fake news cases of multimodal tweets. The right picture has irrelevant images and factual errors, while the left picture has inconsistent text and images.

To mitigate this, subsequent work introduces knowledge-enhanced models, which retrieve or embed facts from large-scale, unimodal Knowledge Graphs (KGs) like WiKiData5M (Wang et al., 2021) to enrich semantic representation (fu et al., 2023; La et al., 2025; Kumar et al., 2025). Some methods use KGs for entity linking and embedding enhancement (Zhao et al., 2022), while others construct auxiliary graphs fused with the news representation graph via GNNs (Wang et al., 2020; Qian et al., 2021). However, this paradigm uses knowledge only as supplementary features for the classifier, inheriting limitations: (i) knowledge loss as relational facts are compressed into holistic vec-

tors; (ii) high computational costs; (iii) limited explainability (Wang et al., 2024; Thomas and Mihailescu, 2025). For example, in Figure 1, methods without external knowledge cannot identify factual errors in (a); even with external knowledge, determining authenticity in (b) based on overall features remains difficult.

Concurrently, Multimodal Knowledge Graphs (MMKGs) have emerged, grounding entities in structured graphs with multimodal features (Zhang et al., 2023; Zhu et al., 2022). Multimodal Knowledge Graph Completion (MKGC), a core task involving entity prediction, relation prediction, and triple classification (Chen et al., 2024), is intrinsically based on structured relational reasoning with good attention to local features, making it suitable for complementing global-feature-based methods.

To fully leverage MKGC's fine-grained capability, we propose MKGRP-FND, a novel framework that reformulates MFND as a relation prediction task within MKGC. Specifically, we develop a universal process that converts news posts into a contextual subgraph and directly assesses authenticity by predicting relationship credibility between entities. This enables fine-grained, fact-based reasoning and provides interpretability by highlighting anomalous subgraph structures.

The main contributions of our work are summarized as follows:

- We firstly propose to transform the multimodal fake news detection task into the relation prediction problem within the MKGC task, and constructed a formulaic process to convert the original news into a predictable graph structure.

- We design a novel model architecture to implement MKGRP-FND method and verify its effectiveness, which consists of a Hybrid Visual-transformer encoder and a dedicated Veracity Evaluation Module (VEM).It achieves comparable performance on the benchmark dataset PHEME.

- We develop a universal interpretability scheme and quantified interpretability indicators, and analyze the interpretability of the MKGRP-FND framework.

2 Related Works

2.1 Multi-modal False News Detection Based on Knowledge Graph

To overcome limitations of traditional detection in identifying factual anomalies, researchers increasingly turn to external knowledge bases, especially KGs. The dominant paradigm uses KGs for semantic enhancement, improving entity and relationship understanding while facilitating modality alignment and feature fusion.

Early approaches focus on simple fact verification. For instance, a knowledge-fusion method (Xie et al., 2023) builds an entity-relationship matrix by querying a knowledge graph, extracts knowledge features via CNN, and fuses them with textual and visual features for classification, directly injecting structured facts.

Recent research constructs more sophisticated graph structures. For example, a framework (Zeng et al., 2025) enhances multimodal fake news detection by building fine-grained knowledge graphs. Another work (Wang et al., 2025) introduces a scene graph attention network that integrates external knowledge and evidence, constructing semantic graphs from text and images.

Despite effectiveness, these methods share limitations: (1) the KGs are mostly unimodal (text-only); visual information is processed separately and fused at feature level, not structurally integrated. (2) KG primarily enhances feature representation and semantic understanding — used for entity disambiguation, semantic embedding enrichment (Cook et al., 2025; Li et al., 2025b), or constructing auxiliary graphs merged with news representation graphs (Wang et al., 2020; Qian et al., 2021). The final decision projects the entire enriched representation into a vector for binary classification. This incurs high computational costs, risks knowledge loss as relational facts are compressed into holistic vectors, and offers limited explainability (Qi et al., 2024).

2.2 Multimodal Knowledge Graph Relationship Prediction

Multimodal Knowledge Graphs extend traditional KGs by associating entities with multimodal features while preserving structured triples. Core challenges include MMKG construction, multimodal representation learning, and Multimodal Knowledge Graph Completion (predicting missing entities or relations).

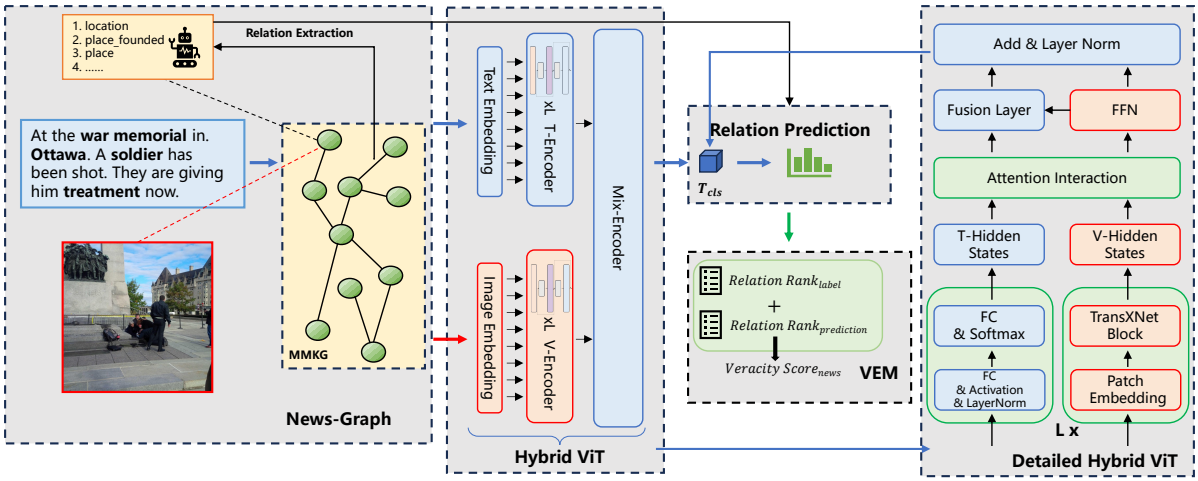

Figure 2: Overall architecture of the MKGRP-FND framework. The framework processes multimodal news through three stages: (1) News-to-Graph Transformation converts text and images into a structured subgraph; (2) Hybrid ViT Encoder performs relation prediction via a TransXNet-based global encoder and a cross-modal M-Encoder; (3) Veracity Evaluation Module (VEM) refines the relation prediction scores using Bayesian optimization to output the final veracity judgement with explainable evidence. The detailed structure of the Hybrid ViT Encoder is expanded on the right side of the figure.

MKGC is fundamentally a relation prediction task. Early methods simply appended multimodal features to entity embeddings. Recent advances focus on deeper modal interaction. A seminal work, MKGformer (Chen et al., 2022), introduces Prefix-Guided Interaction and Correlation-Aware Fusion modules to reduce modal heterogeneity and noise. Building on this, MCCI (Zhang et al., 2025) (2025) emphasizes early interaction between textual and visual features and employs a dynamic relationship-guided decision fusion module.

These works highlight that MMKGC centers on relational reasoning over heterogeneous graphs, unlike feature-enhancement approaches in knowledge-based fake news detection. We modify the MKGC relation prediction paradigm and apply it to our problem: instead of using unimodal KGs to enhance classifier features, we construct MMKG subgraphs for news and directly transform authenticity assessment into relationship prediction. This enables fine-grained consistency checks at entity and relation levels, providing a more direct, structured, and interpretable detection method.

3 Methodology

The MKGRP-FND framework is designed to reframe false news detection as a relation prediction task within a multimodal knowledge graph. The overall architecture, depicted in Figure 2, consists of three core components: (1) the News-to-Graph Transformation module, which converts a raw multimodal news item into a structured, contextual subgraph; (2) the Relation Prediction module, which encodes this subgraph MMKG and performs the core relation prediction; and (3) the Veracity Evaluation Module (VEM), which refines the raw prediction scores into a final, robust veracity judgement. This pipeline enables granular, fact-level reasoning over news content.

3.1 Problem Formulation

The MKGRP-FND framework formalizes multimodal false news detection as a relation prediction task within a Multimodal Knowledge Graph. We will describe the definitions of the two tasks, and then elaborate in detail on how our method accomplishes the transformation.

Multimodal False News Detection Given a multimodal news item $N = (T, I)$, where T represents textual content and I represents the associated image(s), the goal is to predict its veracity label $y \in \{0, 1\}$, where $y = 1$ indicates false news and $y = 0$ indicates true information.

MKGC Relation Prediction Let $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$ denote a Multimodal Knowledge Graph, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relation types, and $\mathcal{F} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of observed factual triples. Each entity $e \in \mathcal{E}$ is associated with multimodal features $\mathbf{x}_e$. The relation prediction task aims to predict the r of a incomplete triple $(h, ?, t) \notin \mathcal{F}$e, typically by calculate the probability of the relation

set.

Paradigm Transformation We propose to bridge these two problems through a structured transformation and reasoning process consisting of three stages: (1) **News-to-Graph Transformation:** From the news item N, we construct a contextual subgraph $\mathcal{G}_N = (\mathcal{E}_N, \mathcal{R}_N, \mathcal{F}_N)$, where $\mathcal{E}_N$ includes extracted textual entities and visual concepts, and $\mathcal{F}_N$ represents their inferred relationships.(2) **Relationship Prediction:** We conduct relationship prediction based on the obtained $\mathcal{G}_N$ with the Hybrid ViT model.(3) **Veracity Evaluation Module:**We defined a mapping ϕ from the probability of the relationship prediction$r_{\text{prediction}}$ to the set of relationship labels of triples $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})$ to the veracity of the news. The scoring function ϕ outputs a probability indicating the likelihood that the news is false. Therefore, the original classification task is reformulated as:$\hat{y}$ = $\mathbb{I}[\phi((\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{F})) > \tau]$, where τ is a decision threshold and $\mathbb{I}[\cdot]$ is the indicator function.

We provide detailed descriptions of these three parts in Sections 3.2, 3.3 and 3.4 respectively.

3.2 News-Graph Transformation Paradigm

The first step is to construct a contextual knowledge subgraph $\mathcal{G}_N = (\mathcal{E}_N, \mathcal{R}_N, \mathcal{F}_N)$, where $\mathcal{E}_N$ includes content T and an associated image I. This process extracts entities and visual concepts as nodes and establishes semantically meaningful relationships as edges.

For textual entity extraction and relation establishment, inspired by KMGCN (Wang et al., 2020), we process the textual corpus of the dataset to build a global co-occurrence statistics matrix. For each news posts text T, we first extract candidate entities (nouns and named entities). The significance of the connection between two entities e_i and e_j within the broader context is quantified using the Pointwise Mutual Information (PMI):

$$\text{PMI}(e_i, e_j) = \log \frac{p(e_i, e_j)}{p(e_i)p(e_j)}, \tag{1}$$

where $p(e_i, e_j)$ is the probability of e_i and e_j co-occurring within a predefined context window, and $p(e_i)$, $p(e_j)$ are their marginal probabilities of occurrence.

An undirected edge $e_{ij} \in E$ is established between e_i and e_j in the subgraph G_{news} if and only if $\text{PMI}(e_i, e_j) > 0$, indicating a statistically significant positive association. Concurrently, we construct a weighted textual adjacency matrix $\mathbf{A}_{\text{text}} \in \mathbb{R}^{n \times n}$ for the entity set:

$$\mathbf{A}_{\text{text}}^{ij} = \begin{cases} \text{PMI}(e_i, e_j), & \text{if PMI}(e_i, e_j) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To assign a semantic label $r_{ij} \in R$ to each edge e_{ij}, we employ an LLM. For each entity pair (e_i, e_j), the LLM is prompted with the pair and their context to generate the top-k most plausible relational phrases. This produces multiple candidate relations, moving beyond a fixed schema and avoiding the need to inflate k in metrics like *HIT@k*. The generated relations are mainly selected from FB15k-237 (a sub-dataset of FB15K (Bordes et al., 2013)) without inverse relations, plus some common verb-noun relations.

For the visual part, we do not specially process full images, as traditional detectors like YOLOv3 are overly detector-dependent and lose global information essential for entity relationship determination. Instead, we address this in Section 3.2's relation prediction.

The final output is a heterogeneous multimodal subgraph $\mathcal{G}_N$, providing a structured, relational representation of key entities, visual content, and their interactions, serving as input for subsequent reasoning.

3.3 Relation Prediction via Hybrid ViT

The core objective of this module is to encode the fused multimodal knowledge graph and predict the possible relations along with their confidence rankings for entity pairs within the graph. The input is the MMKG subgraph $\mathcal{G}_N$. The output is a probability distribution $P(r \mid e_i, e_j, \mathcal{G}_N)$ over potential relations for each entity pair (e_i, e_j) of interest, which serves as the direct input for the subsequent Veracity Evaluation Module.

To achieve this, we design a Hybrid Vision-Transformer (ViT) Encoder, whose architecture is illustrated in Figure 2. It operates in a coordinated manner to perform end-to-end graph representation learning and relational reasoning.

To capture the complex, long-range structural dependencies within the augmented graph, we draw inspiration from the multi-scale hierarchical attention mechanism of TransXNet (Lou et al., 2025), employing it as a lightweight prefix encoder. The purpose of this module is not to perform complex modal fusion directly, but to first generate a global structural representation for each node that is rich in multi-hop neighborhood information.

Specifically, given the initial node feature matrix $\mathbf{X} \in \mathbb{R}^{|V| \times d}$ and the normalized adjacency matrix $\tilde{\mathbf{A}}$ of the graph, we simulate information diffusion at different ranges by computing powers of the adjacency matrix $\tilde{\mathbf{A}}^s$ for scales $s = 1, 2, \ldots, S$. The node representation at each scale s is obtained via a learnable transformation:

$$\mathbf{H}^{(s)} = \tilde{\mathbf{A}}^s \mathbf{X} \mathbf{W}^{(s)}, \tag{3}$$

where $\mathbf{W}^{(s)}$ is a scale-specific weight matrix. Subsequently, an attention gating mechanism adaptively aggregates information from all scales:

$$\mathbf{H}_{\mathrm{g}} = \sum_{s=1}^{S} \mathrm{softmax}_s \left(\mathbf{q}^\top \tanh \left(\mathbf{V} \mathbf{H}^{(s)} \mathbf{u} \right) \right) \cdot \mathbf{H}^{(s)}. \tag{4}$$

Here, $\mathbf{q}, \mathbf{u}, \mathbf{V}$ are learnable parameters. The output $\mathbf{H}_{\mathrm{g}}$ serves as a global contextual prefix, encapsulating the macroscopic topological information of the graph and providing an important structural prior for subsequent fine-grained semantic alignment and relational reasoning.

To bridge the semantic gap between textual entities and visual concepts, we adopt a Cross-Modal Semantic Aligner (M-Encoder) from the modal encoder of MKGformer (Chen et al., 2022). This module achieves deep fusion and alignment of textual and visual node features in a shared semantic space.

First, textual features $\mathbf{H}^t$ (from pre-trained BERT) and visual features $\mathbf{H}^v$ (from TransXNet) are projected into a shared space via separate linear layers. Then, the M-Encoder applies cross-attention, allowing each textual node to attend to all visual nodes and vice versa. This aligns visual concepts with relevant textual descriptions and disambiguates textual entities using visual context. The aligned features $\tilde{\mathbf{H}}$ form the semantic foundation for relation prediction.

Next, we concatenate the global structural information $\mathbf{H}_{\mathrm{g}}$ with the aligned semantic features $\tilde{\mathbf{H}}$, and feed them into a standard multi-layer Transformer encoder to model deep interactions via self-attention. Finally, we extract context-aware node encodings $\mathbf{z}_i$ for each node.

For any entity pair (e_i, e_j) to be predicted, we use their corresponding node encodings $\mathbf{z}_i$ and $\mathbf{z}_j$ to compute a score for each relation $r \in \mathcal{R}$ via a scoring function based on the DistMult factorization:

$$\phi(e_i, r, e_j) = \mathbf{z}_i^\top \, \mathrm{diag}(\mathbf{W}_r) \, \mathbf{z}_j, \tag{5}$$

where $\mathbf{W}_r$ is a learnable diagonal matrix for relation r. After computing scores for all relations, we normalize them via the softmax function to obtain a probability distribution:

$$P(r \mid e_i, e_j) = \frac{\exp\left(\phi(e_i, r, e_j)\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\phi(e_i, r', e_j)\right)}. \tag{6}$$

Finally, we sort all relations in descending order of $P(r \mid e_i, e_j)$ to obtain the predicted relation ranking list $R_{\text{pred}}^{(i,j)}$ for that entity pair. The predicted ranking lists for all entity pairs are then passed as input to the final Veracity Evaluation Module for decision.

In order to train the Hybrid ViT, we design the total loss for fine-tuning the Hybrid ViT with LoRA (Hu et al., 2022) as a standard cross-entropy loss over the relation probability distribution for each entity pair (e_i, e_j):

$$\mathcal{L}_{\text{relation}} = -\frac{1}{|\mathcal{P}|} \sum_{(e_i, e_j) \in \mathcal{P}} \log P(r_{ij}^{\text{gt}} \mid e_i, e_j) \tag{7}$$

where r_{ij}^{gt} is the ground-truth relation label for the pair (i.e., the top candidate from the LLM generation), and $\mathcal{P}$ is the set of all predicted entity pairs.

3.4 Veracity Evaluation Module

The Relation Prediction module outputs a probability distribution over candidate relations for each entity pair in the news subgraph. However, the raw prediction scores may not directly translate to an accurate veracity judgement, especially when the relation prediction performance is moderate. To address this, we design a specialized Veracity Evaluation Module(VEM) that operates on the relation prediction results through a heuristic algorithm and incorporates structural information from the graph construction phase.

For each entity pair (e_i, e_j) in the news subgraph, let R_{pred} be the ranked list of all possible relations sorted by their prediction probabilities in descending order. Denote r_{pred} as the top-ranked predicted relation with the highest probability), and let rank$r$ be its global rank in $Rpred$ (starting from 1).

During the graph construction phase in Section 3.2, for each entity pair, we used an LLM to generate a candidate set of K_0 plausible relations, denoted as $C_{ij} = r_1, r_2, \ldots, r_{K_0}$, sorted by their likelihood. Let rank$k$ be the rank of $rpred$ within C_{ij} if $r_{pred} \in C_{ij}$; otherwise, rank$_k = \infty$.

Dataset	Modality	Metric	Bert	EANN	TextGCN	SAFE	KMGCN	KDCN	MKGRP-FND
PHEME	modal incomplete	Pre.	0.797	0.744	0.792	0.765	0.829	0.836	**0.851**
		Rec.	0.742	0.713	0.803	0.663	0.821	0.827	**0.857**
		F1	0.769	0.728	0.797	0.710	0.825	0.831	**0.854**
	modal complete	Pre.	0.794	0.692	0.741	0.653	**0.876**	0.833	0.841
		Rec.	0.713	0.678	0.628	0.529	**0.877**	0.831	0.826
		F1	0.751	0.685	0.680	0.584	**0.876**	0.832	0.833
Twitter	modal incomplete	Pre.	0.835	0.726	0.798	0.782	0.849	**0.893**	0.872
		Rec.	0.776	0.698	0.786	0.691	0.831	**0.871**	0.865
		F1	0.804	0.712	0.792	0.734	0.840	**0.882**	0.868
	modal complete	Pre.	0.810	0.684	0.743	0.721	0.847	**0.907**	0.904
		Rec.	0.729	0.665	0.637	0.616	0.812	0.861	**0.893**
		F1	0.767	0.674	0.686	0.664	0.829	**0.883**	0.898
Weibo	modal incomplete	Pre.	0.907	0.814	0.863	0.891	0.892	**0.924**	0.912
		Rec.	0.896	0.813	0.851	0.883	0.890	**0.923**	0.906
		F1	0.901	0.813	0.857	0.887	0.891	**0.923**	0.909
	modal complete	Pre.	0.864	0.792	0.825	0.869	0.876	**0.921**	0.924
		Rec.	0.857	0.774	0.816	0.858	0.874	**0.913**	0.917
		F1	0.860	0.783	0.820	0.863	0.875	**0.917**	0.920

Table 1: Performance comparison of different false news detection methods on the modality-complete and modality-incomplete subsets of PHEME. The best results are in **bold** while the second and third best are in underlined."Pre." and "Rec." represent precision and recall respectively.

	modal complete			modal incomplete			inference delay
	Pre.	Rec.	F1	Pre.	Rec.	F1	
MKGRP-FND (our)	0.861	0.857	0.859	0.841	0.826	0.833	1
w/o pre-train(TransXnet)	0.845	0.831	0.838	–	–	–	–
w/o Space Search Module	0.826	0.819	0.822	0.830	0.825	0.827	0.849
w/o TransXnet block	0.804	0.793	0.798	–	–	–	0.962
w/o Attention Interaction	0.728	0.716	0.722	–	–	–	0.775

Table 2: Ablation study on the key components of the MKGRP-FND framework.

We define a relation-specific score Score_{ij} that integrates information from both the prediction ranking and the candidate set ranking:

$$\text{Score}_{ij} = \frac{\left(\frac{a}{\text{rank}_r} + \frac{b}{\text{rank}_k} + c \cdot \mathcal{G}(\text{rank}_r, \text{rank}_k)\right)}{d} \tag{8}$$

where a, b, c are learnable weighting coefficients, $d = 3$ is a normalization factor (the number of terms), $\mathcal{G}(\cdot)$ is a Gaussian similarity function that encourages consistency between the two rankings:

$$\mathcal{G}(\text{rank}_r, \text{rank}_k) = \exp\left(-\frac{(\text{rank}_r - \text{rank}_k)^2}{2\sigma^2}\right), \tag{9}$$

where σ being a scaling parameter, empirically set to 2. This term addresses cases where the model is highly confident about a relation that the LLM deems implausible, which often occurs when the image does not contain relevant textual entities.

The final veracity decision integrates scores from all n entity pairs in the news subgraph. We first normalize the PMI values from Equation 1 associated with each entity pair to the range $[0, 1]$, denoted as $\widetilde{\text{PMI}}_{ij}$.

A news-level veracity score S_{news} is computed as a weighted combination:

$$S_{\text{news}} = \frac{[q \cdot Score' + (1-q) \cdot \sum_{i=1}^{n} \text{Score}_{ij}]}{n}, \tag{10}$$

$$Score' = \sum_{i=1}^{n} \left(\text{Score}_{ij} \cdot \widetilde{\text{PMI}}_{ij}\right) \tag{11}$$

where $q \in [0, 1]$ is a balancing hyperparameter that controls the influence of the PMI-weighted structural information versus the raw relation scores.

The veracity of the news item is then determined by comparing the news-level score S_{news} to a threshold k: if $S_{\text{news}} > k$, the news is classified as true; otherwise, it is classified as false.

In parallel, we validate each individual relation prediction by comparing the relation-specific score Score_{ij} to a separate threshold p: if $\text{Score}_{ij} > p$,

the relation is considered valid; otherwise, it is considered invalid. This provides an additional layer of interpretability by identifying which specific relations contributed to the overall veracity judgement.

The hyperparameters $\Theta = a, b, c, q, k$ critically influence the module's performance. Rather than setting them manually, we employ Bayesian Optimization to search for the optimal parameter configuration that maximizes news veracity judgement accuracy on a validation set.We define the objective function as the accuracy of news-level veracity predictions.

The VEM is particularly effective when the relation prediction performance is limited, as it finds parameter settings that maximize the downstream veracity judgement accuracy without requiring perfect relation predictions.

4 Experiments

4.1 Dataset and Preprocessing

We conduct our experiments on the PHEME, Twitter and Weibo dataset, a widely used benchmark for rumor detection that contains Twitter threads associated with real-world breaking news events. Each instance includes the source tweet, its textual replies, and associated images when available. The dataset provides veracity labels. Inspired by the data processing methodology of KDCN (Sun et al., 2023), we acknowledge that real-world social media posts often suffer from modality incompleteness. To rigorously evaluate the robustness of our multimodal framework, we construct two distinct experimental subsets from PHEME, Modality-Complete Subset ($\mathcal{D}_{mc}$) contains news items that contain both substantive textual content and at least one associated image that is contextually relevant and Modality-Incomplete Subset ($\mathcal{D}_{ic}$) contains all instances from the full PHEME dataset, including both modality-complete items and text-only items.

For both subsets, we adopt a chronological split based on the tweet timestamp to prevent temporal data leakage. Following a common practice, we partition each subset into training, validation, and test sets at 8:1:1. The validation set is used for hyperparameter tuning and early stopping, and the test set is held out for final performance reporting.

4.2 Experimental Setup

Baselines We compare MKGRP-FND against a diverse set of state-of-the-art and representative baselines to comprehensively evaluate its performance, including BERT (Devlin et al., 2019), EANN (Wang et al., 2018), TextGCN (Yao et al., 2019), SAFE (Zhou et al., 2020), KMGCN (Wang et al., 2020), KDCN (Sun et al., 2023).All baseline models are reproduced using their officially released code or our careful re-implementation under the same dataset and split settings to ensure a fair comparison.

Implementation Details For the News-to-Graph Transformation module, the candidate relations for each entity pair are generated using the Qwen2.5-8B-Instruct language model. The implement of Hybrid ViT is inspired by (Chen et al., 2022; Lou et al., 2025). The TransXNet prefix uses a 2-layer GGCN and 3 diffusion scales ($S = 3$). The M-Encoder and the main Transformer stack both consist of 4 layers, with 8 attention heads and a hidden dimension of 768. The visual features are extracted using a CLIP-ViT-B/32 image encoder, and textual entity features are initialized from the final hidden states of a pre-trained `bert-base-uncased` model.We use the AdamW optimizer with a learning rate of 2×10^{-5} and a linear warmup schedule over the first 10% of steps. The model is trained with a batch size of 32 for a maximum of 50 epochs, with early stopping based on the validation loss (patience=10). The hyperparameter set $\Theta = \{a, b, c, q, k, p\}$ of VEM is optimized via Bayesian search using the `scikit-optimize` library over 50 iterations on the validation set of $\mathcal{D}_{mc}$. The search ranges are: $a, b, c \in [0, 3]$, $q \in [0, 1]$, $k \in [0, 1]$, $p \in [0, 1]$.

4.3 Results

Table 1 summarizes the main results, which are the median values obtained from multiple repeated experiments. On the modality-incomplete subset ($\mathcal{D}_{ic}$), MKGRP-FND achieves the best F1 on PHEME and ranks second on Twitter and Weibo, demonstrating strong robustness. On the modality-complete subset ($\mathcal{D}_{mc}$), it remains highly competitive, ranking among the top two on all datasets, and even slightly surpasses KDCN on Weibo. These results validate that MKGRP-FND excels under realistic incomplete multimodal conditions while staying highly competitive under ideal settings.

4.4 Ablation Study

We perform ablation experiments to assess each core component (Table 2), also reporting relative inference delay. Removing the pre-trained TransXNet prefix or the entire TransXNet block

Qwen-7B	Qwen-7B	Qwen-72B	Manual	Llama-80B	GPT4	Manual
post	3.571	3.372	3.224	3.453	3.325	3.017
post + prediction	4.468	4.749	4.451	4.407	4.618	4.529
MAE	0.182	0.223	–	0.279	0.199	–
MAE(between explanation)	0.897	1.377	1.227	0.954	1.293	1.512

Table 3: Results of quantitative interpretability analysis using different LLM judges, selected 500 samples from each of PHEME and Twitter. Qwen refers to the Qwen2.5-VL model, and post represents those were correctly predicted by our method.

leads to a noticeable performance drop while reducing delay only marginally, indicating that the multi-scale global structural prior provides useful context with little cost. Replacing the Bayesian-optimized VEM with a simple PMI-weighted voting method causes a clear performance decline at a somewhat reduced delay, confirming a worthwhile accuracy–efficiency trade-off. Removing the cross-modal Attention Interaction results in a severe performance loss with moderate delay reduction, demonstrating that deep semantic alignment between modalities is essential for effective relation prediction. Overall, the full model achieves the best performance, and the accuracy gains justify the moderate computational overhead.

4.5 Interpretability Analysis

Beyond detection performance, a key advantage of MKGRP-FND is its inherent ability to provide structured, evidence-based explanations for its verdicts. To quantitatively evaluate this interpretability, we design a novel assessment methodology that leverages both state-of-the-art LLMs and LVLMs as judges, complemented by human evaluation.

Evaluation Protocol. We randomly sample 500 news instances from each of the PHEME and Twitter test sets, totaling 1,000 samples. For each instance, we extract the set of predicted entity-relation triples from our model. We then construct two types of prompts for the LLM/LVLM judges: (1) **Baseline Prompt**: Provides the news text, associated image, and the ground-truth veracity label, asking the model to generate a reasoning explanation. (2) **Enhanced Prompt**: In addition to the above, it provides the key predicted relations identified by MKGRP-FND, and asks for an explanation.

The quality of the generated explanations is automatically scored by separate, powerful LLMs and LVLMs on a 5-point scale based on clarity, logical consistency, and grounding in the provided evidence. To ensure fairness and avoid model bias, we also conduct **human evaluation**: all generated explanations are mixed and anonymized, then rated by three independent annotators following the same criteria. To measure the specific contribution of our model's relational predictions to explainability, we calculate Mean Absolute Error (MAE) as the average absolute difference in scores for the same instance under the two prompt conditions. We also compute MAE between different judges under the same prompt condition to assess inter-judge variance.

Results (Table 3) show a consistent and significant explanation gain across all judges when our relational predictions are provided; the enhanced prompt always yields higher scores, confirming that structured relations improve explanation quality.

Notably, MAE(between explanation) is substantially larger than MAE(between model). This indicates that the variance due to adding our relational evidence far exceeds the inherent variance between different judges, providing strong evidence that our predicted relational triples are meaningful, high-impact explanatory elements. Thus, MKGRP-FND not only detects false news but also reveals the semantic and relational discrepancies behind its conclusions.

5 Conclusion

In this study, we introduce MKGRP-FND, a novel framework that demonstrates remarkable effectiveness by reframing MFND as a relation prediction task within MKGC. It enables fact-level, interpretable reasoning by transforming news into a graph structure and assessing the plausibility of its relational claims. Through comprehensive experiments on public benchmark, it delivers competitive, near state-of-the-art performance. Furthermore, quantitative analysis and case studys confirms that MKGRP-FND provides structured, evidence-based explanations for its decisions, significantly enhancing model transparency and interpretability.

Limitations

Despite the promising results achieved by MKGRP-FND, several limitations should be acknowledged.

First, our framework is mainly designed for Chinese and English examples. In practice, for emerging events or niche domains, the external KG may lack relevant facts, which could hinder the link prediction process and lead to false negatives or unreliable explanations.

Second, the computational cost of the Hybrid ViT encoder, particularly the TranXNet prefix and cross-modal attention modules, is higher than simpler baselines like BERT or TextGCN. Although our ablation study shows the performance gains justify the overhead, real-time or resource-constrained applications may still find the inference latency challenging when processing large volumes of social media data.

Ethics

In this study, we invited three volunteers to annotate the data and obtain the average human results. Each volunteer was paid RMB 1 per 10 samples. The datasets and models used in this study are from open access repositories and strictly follow their open source license agreements. These sources are only used for research purposes and do not violate their open source terms. This study uses the API provided by OpenAI and some open source models to verify the interpretability and the polishing of the text.

References

Ifeoma Adaji and Linda O. Okpanachi. 2024. Detecting fake news spreaders on social media using posts content versus profile information. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6190–6198.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.

Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhuo Chen, Yichi Zhang, Yin Fang, Yuxia Geng, Lingbing Guo, Xiang Chen, Qian Li, Wen Zhang, Jiaoyan Chen, Yushan Zhu, Jiaqi Li, Xiaoze Liu, Jeff Z. Pan, Ningyu Zhang, and Hua zeng Chen. 2024. Knowledge graphs meet multi-modal learning: A comprehensive survey. *ArXiv*, abs/2402.05391.

Owen Cook, Charlie Grimshaw, Ben Peng Wu, Sophie Dillon, Jack Hicks, Luke Jones, Thomas Smith, Matyas Szert, and Xingyi Song. 2025. Efficient annotator reliability assessment and sample weighting for knowledge-based misinformation detection on social media. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3348–3358, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifang fu, Huanxin Peng, and Shuai Liu. 2023. Kgmfend: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing*, 79(16):18417–18444.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2023. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *AAAI Conference on Artificial Intelligence*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Echofaked: improving fake news detection in social media with an efficient deep neural network. *Neural Computing and Applications*, 33(14):8597–8613.

Anuj Kumar, Pardeep Kumar, Abhishek Yadav, Satyadev Ahlawat, and Yamuna Prasad. 2025. KG-FakeNet: A knowledge graph-enhanced model for fake news detection. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 109–122, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Tuan-Vinh La, Minh-Hieu Nguyen, and Minh-Son Dao. 2025. Kgalign: Joint semantic-structural knowledge encoding for multimodal fake news detection. *ArXiv*, abs/2505.14714.

Xuefeng Li, Jian Wei, Chensu Zhao, Xiaqiong Fan, and Yuhang Wang. 2025a. Multi-domain fake news detection method based on generative adversarial network and graph network. *Know.-Based Syst.*, 319(C).

Yiheng Li, Yang Yang, Zichang Tan, Huan Liu, Weihua Chen, Xu Zhou, and Zhengdong Lei. 2025b. Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252.

Meng Lou, Shu Zhang, Hong-Yu Zhou, Sibei Yang, Chuan Wu, and Yizhou Yu. 2025. Transxnet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11534–11547.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. *Proceedings of the 29th ACM International Conference on Multimedia*.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13052–13062.

Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(3).

Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and Philip S. Yu. 2023. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12736–12749.

Jonathan John Thomas and Radu-Casian Mihailescu. 2025. An explainable kg-rag-based approach to evidence-based fake news detection using llms. *Proceedings of the AAAI Symposium Series*, 6(1):152–154.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2452–2463, New York, NY, USA. Association for Computing Machinery.

Jinghong Wang, Hongbo Yang, Xizhao Wang, Wei Wang, and Yanan Li. 2025. Magnet: A multimodal knowledge-augmented graph network for early-stage misinformation detection. *Neurocomputing*, 658:131533.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 540–547, New York, NY, USA. Association for Computing Machinery.

Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2024. Knowledge graph enhanced heterogeneous graph neural network for fake news detection. *IEEE Transactions on Consumer Electronics*, 70(1):2826–2837.

Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, Shan Xue, and Hao Fan. 2023. Heterogeneous graph neural network via knowledge relations for fake news detection. In *Proceedings of the 35th International Conference on Scientific and Statistical Database Management*, SSDBM '23, New York, NY, USA. Association for Computing Machinery.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239, Toronto, Canada. Association for Computational Linguistics.

Yirong Zeng, Xiao Ding, Bibo Cai, Ting Liu, and Bing Qin. 2025. Exploring large language models for effective rumor detection on social media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2552, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiquan Zhang, Jianwu Dang, Yangping Wang, and Shuyang Li. 2025. Mcci: A multi-channel collaborative interaction framework for multimodal knowledge graph completion. *Information Processing & Management*, 62(4):104156.

Yuchen Zhang, Xing Su, Jia Wu, Jian Yang, Hao Fan, and Xiaochuan Zheng. 2023. Emoknow: Emotion- and knowledge-oriented model for covid-19 fake news detection. In *Advanced Data Mining and Applications*, pages 352–367, Cham. Springer Nature Switzerland.

Xiang Zhao, Weixin Zeng, Jiuyang Tang, Wei Wang, and Fabian M. Suchanek. 2022. An experimental study of state-of-the-art entity alignment approaches. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2610–2625.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Similarity-aware multi-modal fake news. *Advances in Knowledge Discovery and Data Mining*, 12085:354 – 367.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36:715–735.