Beyond Text: Fine-Grained Multi-Modal Fact Verification with Hypergraph Transformers

Hui Pang[1], Chaozhuo Li[1]*, Litian Zhang[2], Senzhang Wang[3], Xi Zhang[1]

[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
[2]School of Cyber Science and Technology, Beihang University, China
[3]School of Computer Science and Engineering, Central South University, China
{hui_pang2022, lichaozhuo, zhangx}@bupt.edu.cn, litianzhang@buaa.edu.cn, szwang@csu.edu.cn

Abstract

Fact verification has become increasingly vital in the internet age, driven by the proliferation of false claims and political misinformation. While traditional methods rely predominantly on text-based evidence, multi-modal evidence introduces richer sources of information, offering valuable insights for claim verification. Existing multi-modal verification models often focus on superficial correlations between claims and evidence, neglecting the complex semantic interactions present in fine-grained multi-modal signals. In this paper, we propose a novel framework for multi-modal fact-checking, named Hypergraph Transformer-based Multi-modal Fact-Checking (HGTMFC). Our approach captures high-order relationships between different modalities of evidence and claims by leveraging hypergraphs. HGTMFC models the intricate relationships among evidence across various modalities and enhances information propagation through a transformer-based mechanism embedded within the hypergraph. Moreover, we utilize linegraphs to refine this propagation process, further strengthening the model's reasoning capabilities. Experiments on benchmark datasets demonstrate that our model significantly outperforms existing approaches in multi-modal fact verification.

Figure 1: An illustration of fine-grained correlations in multi-modal checking tasks.

Introduction

With the rapid advancement of social networks, there has been a significant increase in the dissemination of false and misleading content online (Allcott and Gentzkow 2017). This situation underscores the urgent need for automated fact-verification systems to preserve the integrity of information dissemination. Fact verification involves the rigorous evaluation of the accuracy and truthfulness of claims through comprehensive investigation and comparison with corroborated evidence, which is crucial for mitigating the spread of false claims, online rumors, and political deception (Guo, Schlichtkrull, and Vlachos 2022).

The nucleus of fact-checking lies in extracting valuable knowledge or clues (i.e., evidence) related to the input claim from a large corpus, such as verified news sources and knowledge graphs (Kim et al. 2023). Traditionally, evidence has been primarily text-based (Zhang et al. 2024a). Recently, with the proliferation of social media, multi-modal content as evidence candidates has become increasingly common. As shown in Figure 1, recent news published on verified platforms regarding an incident involving Donald Trump being shot typically includes both images and text, which might be integrated into the evidence candidate set. Consequently, the task of retrieving and verifying textual claims using multi-modal evidence, i.e., multi-modal fact-checking, has attracted significant attention.

Existing research on multi-modal fact-checking has demonstrated substantial progress in the extraction of evidential information by complementing missing semantics with visual evidence (Yao et al. 2023; Abdelnabi et al. 2022; Du et al. 2023; Zhang et al. 2023b). For instance, Yao et al. (Yao et al. 2023) and Abdelnabi et al. (Abdelnabi et al. 2022) initially perform single-modal fact-checking predictions using neural networks, subsequently applying multi-modal fusion techniques. This involves extracting vectors from each modality and integrating them into a unified space for evi-

*Corresponding author: Chaozhuo Li.

dence verification tasks. Du et al. (Du et al. 2023) enhance large-scale base models with adapters and employ multi-modal multi-type fusion modules to elucidate relationships across modalities and between different types of evidence, such as statements and documents.

Despite the promising performance of existing models, they predominantly emphasize superficial correlations between claims and evidence. Typically, these models separately encode claims and evidence as distributed representations, which are subsequently processed through a classification layer to reach a final decision (Du et al. 2023; Li et al. 2018; Zhang et al. 2024c). This coarse-grained paradigm overlooks the intricate semantic interactions inherent in fine-grained multi-modal signals. As shown in Figure 1, the complex cross-domain interactions—such as token-token, token-patch, or patch-patch relationships within multi-modal evidence—as well as the semantic interplay between textual entities and visual objects, are crucial for successful claim verification (Zhang et al. 2024b). Neglecting these fine-grained interactions can lead to two significant issues: (1) insufficient integration of evidence across modalities, which may result in an over-reliance on a single modality and the neglect of information from others, and (2) difficulty in modeling higher-order relationships essential for comprehensive multi-modal fact-checking. These relationships encompass both intra-modal and inter-modal information transmission.

Capturing intricate fine-grained cross-modal signals within a unified framework remains a formidable challenge. A prevalent approach involves modeling the informative components of multi-modal data as nodes within a graph, with their interrelationships depicted by edges (Tao et al. 2020; Yin et al. 2020; Li et al. 2021). Graph neural networks (GNNs) are subsequently utilized to integrate this cross-domain information. However, conventional GNNs encounter limitations in capturing higher-order features, as they predominantly focus on node feature aggregation and often neglect the rich information contained in the connecting edges (Zhang et al. 2023a; Li et al. 2017). Moreover, GNNs typically function through a one-to-one mapping process, which poses additional difficulties in modeling data correlations that require nuanced information fusion, such as the relationships between phrases in textual claims and evidence or between patches/objects in image evidence. In such scenarios, establishing connections between distant nodes necessitates multiple hops, relying on multi-layer aggregation within the GNN. Unfortunately, the challenges of over-smoothing and noise, particularly in deep GNN architectures, can adversely affect the model's overall performance (Dong and Kluger 2023; Chen et al. 2020).

To address the aforementioned challenges, we propose a novel Multi-modal Fact-Checking framework, Hypergraph Transformer-based Multi-modal Fact-Checking (HGTMFC), designed to model the high-order relationships between different modalities of evidence and claims. Unlike traditional multi-modal fact-checking approaches, HGTMFC jointly models the high-order relationships among evidence across various modalities via hypergraphs. HGTMFC begins with retrieving relevant evidence related to a given claim from both text and image corpora. Features from both the claim and the retrieved evidence, including textual and visual elements, are extracted to construct a hypergraph that encompasses text evidence nodes, claim nodes, and image evidence nodes. By leveraging the transformer mechanism for hypergraph information propagation, this approach facilitates effective information flow within the hypergraph structure, enhancing the integration of high-order information between multi-modal claims and evidence. To further capture the complex relationships between claims and different evidence hyperedges, a line graph is employed to optimize the information propagation process, thereby improving reasoning and comprehension of evidence. Finally, a supervised loss function is applied for evaluation and model optimization.

The main contributions are summarized as follows:

- We propose a novel hypergraph-based model HGTMFC designed to capture fine-grained modality correlations, thereby facilitating multi-modal fact verification.
- We integrate the group transformer mechanism and line-graph information propagation into the hypergraph neural network, enabling enhanced evidence reasoning.
- Experimental validation on benchmark datasets shows that our model effectively fuses evidence and claim information, surpassing state-of-the-art methods and achieving superior performance.

Problem Definition

Given the input claim c, a set of retrieved text evidence $E_t = \{e_{t_1}, e_{t_2}, ..., e_{t_{|E_t|}}\}$, and a set of retrieved visual evidence $E_v = \{e_{v_1}, e_{v_2}, ..., e_{v_{|E_v|}}\}$, multi-modal fact checking aims to classify the claim c into one of the categories $\hat{y}_i \in \{\mathrm{SUP}, \mathrm{REF}, \mathrm{NEI}\}$ (i.e., Supported, Refuted or Not Enough Info). The goal is to predict the truthfulness of the claim based solely on the information derived from both the textual and visual evidence.

Methodology

To simplify the presentation, this paper focuses on text and image modalities though the proposed model, **HGTMFC**, is extendable to incorporate audio and video data. The framework is depicted in Figure 2. Initially, relevant text and image evidence pertaining to a claim is retrieved. This is followed by feature extraction from both the claim and the evidence. A hypergraph is then constructed, encompassing nodes for the claim, text, and images. Multiscale features are employed to extract both fine- and coarse-grained information. The propagation of information within the hypergraph is integrated with transformers, and line graphs are used to extract additional inter-edge information to enhance reasoning capabilities. Finally, a supervised loss function is applied for model evaluation and optimization.

Multi-modal Evidence Retrieval

In this subsection, we detail the evidence retrieval component of our framework. For text evidence retrieval, we utilize the SBERT model (Reimers and Gurevych 2019), which computes semantic similarity between the claim and

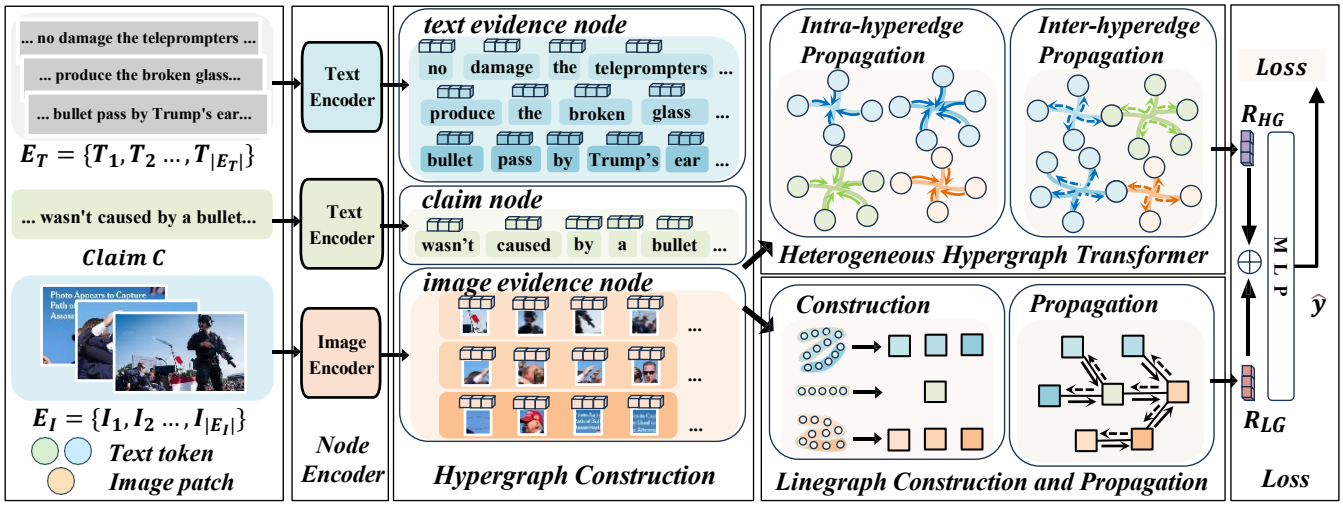


Figure 2: The overview framework of HGTMFC model.

sentences within documents. Candidate evidence is initially ranked and subsequently re-ranked by SBERT (Reimers and Gurevych 2019) based on cosine similarity to select the most pertinent textual evidence. In the case of image retrieval, the CLIP model (Radford et al. 2021) generates feature representations for both the claim and images, identifying relevant images through cosine similarity. This methodology effectively retrieves multi-modal evidence, thereby supporting the fact-checking process.

Multi-modal Node Encoder

The constructed hypergraph includes three node types: tokens from claims, tokens from textual evidence, and patches from visual evidence. For text, each token is treated as a node, along with entities extracted using Spacy. For images, patches are generated at various resolutions and perspectives. A multi-modal node encoder converts these data types into distributed representations. For textual evidence and claims, tokens are processed through CLIP's text encoder, which uses a Transformer to convert sequences into token embeddings. Let $T = \{t_1, t_2, \ldots, t_n\}$ represent a token sequence, and the CLIP model transforms it into embeddings $X_t = \text{CLIP}_{\text{Text Encoder}}(T) = \{\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_n}\}$. For image evidence, the CLIP model uses a CNN to segment images into regions (patches), each encoded as an embedding $\mathbf{p_j}$. Let $I = \{p_1, p_2, \ldots, p_m\}$ denote patches, which are transformed into embeddings $X_p = \text{CLIP}_{\text{Image Encoder}}(I) = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_m}\}$. These embeddings are then used for multi-modal information fusion in fact-checking.

Modality Hypergraph Construction

In this section, we present the hypergraph construction methodology. Our model conceptualizes each token from text and each patch from images as nodes within the hypergraph, with embedding vectors generated by the CLIP model. The modality hypergraph is constructed with three types of hyperedges: text evidence hyperedges, claim hyperedges, and image evidence hyperedges. As shown in Figure 2, these hyperedges are represented by blue, green, and yellow lines, respectively, each connecting multiple vertices corresponding to its modality. For text evidence, each sentence is represented by a hyperedge that connects all token nodes. For instance, the i-th text evidence hyperedge T_i connects all token nodes $\{t_{i1}, t_{i2}, \ldots, t_{in}\}$, forming the set $\text{Edge}_{T_i} = \{t_{i1}, t_{i2}, \ldots, t_{in}\}$. Similarly, the i-th image hyperedge I_i connects all patch nodes $\{p_{i1}, p_{i2}, \ldots, p_{in}\}$, forming the set $\text{Edge}_{I_i} = \{p_{i1}, p_{i2}, \ldots, p_{in}\}$. The incidence matrix for claim and evidence between the vertex set V and the hyperedge set E is defined as:

$$\mathbf{H}_m(i,j) = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{if } v_i \notin e_j. \end{cases} \quad (1)$$

This construction effectively encapsulates the complex multi-modal relationships by modeling text and image data within a unified hypergraph framework.

Multi-modal Hypergraph Group Transformer

Upon constructing the hypergraph, we introduce a novel multi-modal hypergraph group transformer that exploits the properties of hyperedges within the model. Traditional hypergraph neural networks (HGNNs) facilitate information propagation from vertices to hyperedges and subsequently back to vertices (Feng et al. 2019). Initially, nodes are encoded into embedding vectors. Leveraging the hypergraph structure, information from vertices associated with a hyperedge is aggregated onto the hyperedge to capture higher-order group characteristics. This aggregated information is then propagated back to the vertices linked to the hyperedge.

However, this fundamental HGNN information propagation paradigm has certain limitations (Zhao et al. 2021). (1) The CLIP model, designed for text and image matching tasks, may produce pre-learned and fixed node semantic vectors. When applied to claim classification tasks, this static encoding may not align well with the dynamic requirements of subsequent tasks, potentially reducing modeling effectiveness. End-to-end co-training is likely to enhance per-

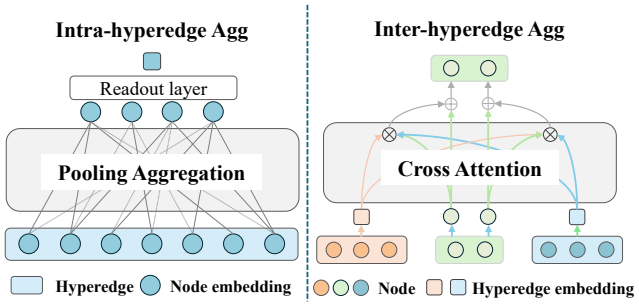


Figure 3: Framework of the two core modules in the multi-modal Hypergraph Group Transformer.

formance by integrating semantic learning with graph topology modeling. (2) During fact verification, the significance of node information can vary within the same modality hyperedges and across different modal hyperedges.

To address these challenges, we propose a hypergraph transformer module designed to enhance the integration of information across multiple modalities. As show in figure 3, this module incorporates semantic encoding and hypergraph propagation, facilitating end-to-end co-training. Specifically, our approach achieves the following: (1) It treats multi-modal semantic understanding parameters as trainable, establishing a co-training paradigm. (2) It effectively captures the significance of various modalities by implementing information propagation from nodes to hyperedges and back to nodes.

Intra-hyperedge Aggregation The module for intra-hyperedge information aggregation aims to learn representations of claim and evidence hyperedges by capturing semantic correlations at different granularities. Given a hyperedge e_i and its connected node set $\mathcal{V}(e_i) = \{v_1, v_2, ...v_m\}$, each node $v \in \mathcal{V}(e)$ is associated with a textual token or an image patch, which is fed into a pretrained multi-modal model to obtain representations. All nodes in $\mathcal{V}(e_i)$ form an embedding set $\mathcal{S} = \{\mathbf{s}_i \in R^d\}_{i=1}^m, \mathcal{S} \in R^{m \times d}$. Inspired by the fundamental Transformer (Vaswani 2017), the multi-modal model is jointly trained with the hyperedge transformation (HT) module. By introducing a learnable reference set $\mathcal{Z} = \{\mathbf{z}_i \in R^d\}_{i=1}^p$, elements from $\mathcal{V}$ are aggregated into fixed-size representations. Each $\mathbf{z}_i$ can be viewed as a pooling unit gathering semantic information from $\mathcal{V}$. The calculation of self-attention among intra-hyperedge nodes can be formulated as:

$$\begin{aligned} \mathbf{Q}^v = \mathbf{W}_S^q \mathbf{S}, \mathbf{K}^v = \mathbf{W}_S^k \mathbf{S}, \mathbf{V}^v = \mathbf{W}_S^v \mathbf{S}, \\ \mathbf{Z} = \text{self\_att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}^v \cdot \mathbf{K}^{v\top}}{\sqrt{d}})\mathbf{V}^v, \end{aligned} \tag{2}$$

where, $\mathbf{W}_S^q, \mathbf{W}_S^k, \mathbf{W}_S^v \in R^{p \times m}$ are trainable parameter matrices. Then the superedge e_i is represented as:

$$\mathbf{e}_i = \mathbf{W}_e[\mathbf{z}_1 : \mathbf{z}_2 : ... : \mathbf{z}_p]^\top, \tag{3}$$

where [:] represents concatenate operation, $\mathbf{W}_e \in R^p$ is the trainable parameter matrix.

However, the basic transformer attention mechanism has certain limitations, as it only accounts for relationships between individual tokens or patches at a single granularity level. To address this, we propose an enhancement to the conventional self-attention mechanism through grouped attention, which captures interactions among neighborhoods of tokens or patches.

Specifically, We associate neighborhoods of tokens or patches and replace certain entries in the Query, Key, and Value matrices with the aggregated representations of entire groups, thereby generating proxy representations for each group. Aggregation can be performed using methods such as max pooling or convolution. Typically, we partition Q, K, and V into n segments, with each segment's feature represented as X_i, where $i \in 1, 2, ..., n$ (with X representing Q, K, or V). The aggregation operation for each segment is denoted as $Agg_i(X_i)$. We then concatenate the aggregated results of all segments to produce X', yielding the group proxies Q', K', and V'. These group proxies are subsequently used for attention computation to generate the final output Z', formulated as:

$$\mathbf{Z}' = \text{self\_att}(\mathbf{Q}', \mathbf{K}', \mathbf{V}') = \text{softmax}(\frac{\mathbf{Q}'^v \cdot \mathbf{K}'^{v\top}}{\sqrt{d}})\mathbf{V}'^v. \tag{4}$$

Given that the input to attention calculation now comprises group proxies, we model the correlations among $K \times K$ tokens, where K denotes the kernel size of the aggregator and may vary across segments. This strategy extends beyond individual tokens or patches, enabling more comprehensive correlation modeling. Consequently, this method allows us to capture relationships between groups of varying sizes, thereby enhancing overall model performance.

Heterogeneous Inter-hyperedge Aggregation The attention mechanism for information propagation between hyperedges is designed to learn the attention weights between vertices and their associated hyperedges. This mechanism recognizes that different vertices have varying degrees of importance with respect to their connected hyperedges, and vice versa. It thereby enhances node representation by aggregating topological information from the connected hyperedges. Therefore, we employ a scaled dot product attention mechanism with a heterogeneous attention approach to propagate information from the hyperedge to the vertex. Specifically, consider the target node v_i with its connected hyperedge set $\mathcal{E}(v_i) = \{e_1, e_2, \ldots, e_t\}$. Each hyperedge $e_i \in \mathcal{E}(v_i)$ has a corresponding learned representation $\mathbf{e_i}$. The attention score between node v_i and its connected hyperedges can thus be expressed as:

$$\begin{aligned} \mathbf{q}_i^v = \mathbf{W}^q \mathbf{v}_i, \mathbf{k}_j^e = \mathbf{W}^k \mathbf{e}_j, \mathbf{v}_j^e = \mathbf{W}^v \mathbf{e}_j, \\ a_{i,j} = \text{att}(\mathbf{v}_i, \mathbf{e}_j) = \text{softmax}_{e_j \in \mathcal{E}_v(v)}(\frac{\mathbf{q}_i^v \cdot \mathbf{k}_j^e}{\sqrt{d}}), \end{aligned} \tag{5}$$

where $\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v \in R^{d \times d}$ are trainable parameter matrices. Then, the representation of each node v is implemented based on a weighted aggregation of the projected

hyperedge representation and the attention score:

$$\tilde{\mathbf{v}}_i = (1-\beta)\mathbf{v}_i + \beta \sum_{e_j \in \mathcal{E}(v)} a_{i,j}\mathbf{v}_j^e. \tag{6}$$

Thus, the hypergraph representation $\mathbf{r}_{HG}$ is:

$$\mathbf{r}_{HG} = (\sum_{i=1}^{|V|} \tilde{\mathbf{v}}_i) \;/\; |\mathcal{V}|. \tag{7}$$

Linegraph Information Propagation

Hypergraphs effectively capture high-order features of claims, text, and image evidence, modeling complex relationships among data elements. However, traditional hypergraph models may fail to fully capture interactions between high-order substructures. To address this, we introduce a line graph, which converts hyperedges into nodes and creates edges based on shared vertices. This approach captures coarse-grained interactions between hyperedges and enables a deeper exploration of the graph's semantic structure.

Given the hypergraph's incidence matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$, we construct the line graph $L(G) = (\mathcal{V}_L, \mathcal{E}_L)$. In this line graph, nodes correspond to hyperedges of the hypergraph, with node embeddings initialized from the hyperedge embeddings $\mathbf{R} \in \mathbb{R}^{|\mathcal{E}|\times d}$. The adjacency matrix $\mathbf{A}_L \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ is defined such that $|\mathcal{E}| = |\mathcal{V}_L|$, indicating that the number of hyperedges in the hypergraph corresponds to the number of nodes in the line graph. We then utilize a Graph Convolutional Network (GNN) to propagate information across the line graph for representation learning, which is computed as:

$$\begin{aligned} \mathbf{G} &= \sigma\left(\tilde{\mathbf{D}}_L^{-\frac{1}{2}}(\mathbf{A}_L + \mathbf{I})\tilde{\mathbf{D}}_L^{-\frac{1}{2}}\mathbf{R}\mathbf{W}\right), \\ \mathbf{r}_{LG} &= \frac{1}{|\mathcal{V}_L|}\sum_{i=1}^{|\mathcal{V}_L|}\mathbf{g}_i, \end{aligned} \tag{8}$$

where $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{|\mathcal{V}_L|}]^\top \in \mathbb{R}^{|\mathcal{V}_L|\times d}$, $\tilde{\mathbf{D}}_L$ is the degree matrix of $\tilde{\mathbf{A}}_L$, $\mathbf{R}$ represents the hyperedge embeddings in the hypergraph, and $\mathbf{W}$ is the trainable parameter matrix. This process yields the representation $\mathbf{R}_{LG}$, providing an alternative perspective on the original data and enhancing the capture of high-order semantic features.

Objective Function

We concatenate the graph representations $\mathbf{r}_{HG}$ and $\mathbf{r}_{LG}$ obtained from two branches, and input this connected representation into a classifier (a Multi-Layer Perceptron, MLP) to predict the label $\hat{y}$. The supervised loss is then calculated using cross-entropy loss function.

$$\begin{aligned} \hat{y} &= [\mathbf{r}_{HG} \;:\; \mathbf{r}_{LG}]\mathbf{W}_{\mathbf{R}}, \\ Loss_s &= -\sum_i y_i \log(\hat{y}_i), \end{aligned} \tag{9}$$

where $[\;:\;]$ denotes the concatenation operation, $\mathbf{W}_{\mathbf{R}} \in \mathbb{R}^{2d\times l_c}$ is a trainable parameter matrix, and l_c represents the number of categories. In this study, $l_c = 3$. y_i denotes the ground truth label, and $\hat{y}_i$ represents the predicted label over the classes for the i-th instance. The cross-entropy loss function quantifies the difference between the predicted probability distribution and the actual label distribution.

Experiment

Experimental Settings

Dataset In our experiments, due to the limited availability of relevant datasets, we selected the Mocheg dataset (Yao et al. 2023) to analyze multi-modal fact-checking. The dataset includes 15,601 claims from PolitiFact and Snopes, labeled for truthfulness and supported by verified ruling statements. Mocheg provides a diverse evidence base with 33,880 textual paragraphs and 12,112 images, making it ideal for multi-modal fact-checking. The dataset is split into training, validation, and test sets, with 11669, 1490 and 2442 samples respectively and labels available for the first two. We used Mocheg to assess the effectiveness of our architecture in this domain.

Baselines Due to the limited research on multi-modal fact verification methods, we choose five text-only methods, two image-only methods and three multi-modal methods.Text only model we select GEAR (Zhou et al. 2019), KGAT (Liu et al. 2019), HESM (Subramanian and Lee 2020), Triple-Check-w text (Du et al. 2023) and MOCHEG-w text (Yao et al. 2023). Image only model we select Triple-Check-w image (Du et al. 2023) and MOCHEG-w image (Yao et al. 2023) Multi-modal model we select Triple-Check (Du et al. 2023), Ino (Zhang et al. 2023b) and MOCHEG (Yao et al. 2023). ‘

Evaluations We evaluate our model based on four traditional evaluation metrics: label accuracy(LA), precision(Pre), recall(Rec) and F1 score(F1). LA measures the classification accuracy for the labels SUPPORTS, REFUTES, and NOT ENOUGH INFO, independent of the retrieved evidence. F1 score is an indicator employed to measure the performance of a binary or multi-class model, which considering the precision and recall of the model.

Implementation details In our study, we implemented evidence retrieval methodologies as described in the Mocheg dataset. For this purpose, we utilized the CLIP model as an encoder for both text and images. The CLIP model's maximum text length was set to 77 tokens, and images were processed after cropping to 16x16 pixels. We employed the Adam optimizer with a learning rate of 2×10^{-6}, a weight decay of 1×10^{-5}, and a batch size of 16. Random node sampling was set to 50, the residual connection weight β was configured at 0.5, the reference set size p was established at 6, and the number of hypergraph propagation layers was set to 2. Early stopping was applied when the validation loss did not decrease within 20 epochs, with a maximum of 50 epochs allowed for training. Our implementation was applied to both retrieved evidence and gold evidence, with all models developed using PyTorch.

Main Results

We evaluate our HGTMFC model against ten baseline methods, including five text-only approaches, two image-only approaches and three multi-modal based approaches. The primary experimental results are summarized in Table 1.

	Model	Retrieval Evi				Gold Evi			
		LA(%)	Pre(%)	Rec(%)	F1(%)	LA(%)	Pre(%)	Rec(%)	F1(%)
Text-only	GEAR	39.31	26.60	38.88	28.99	41.89	28.79	41.43	31.14
	KGAT	44.76	46.17	44.43	40.41	48.65	51.30	48.31	44.45
	HESM	45.82	46.20	45.60	43.80	48.73	49.70	48.50	46.80
	Triple-Check-w text	44.19	44.61	43.97	42.22	46.97	48.00	46.75	45.14
	MOCHEG-w text	44.59	45.40	44.38	42.80	49.47	50.65	49.24	47.59
	HGTMFC-w text	46.68	47.14	46.45	44.66	50.74	52.04	50.51	48.83
Image-only	Triple-Check-w image	43.28	43.63	43.07	41.35	45.58	46.31	45.36	43.69
	MOCHEG-w image	41.73	42.25	41.53	39.91	47.01	47.87	46.79	45.12
	HGTMFC-w image	44.76	45.51	44.55	42.92	48.89	50.12	48.67	47.06
Multi-modal	Triple-Check	45.33	45.64	45.11	43.27	49.06	50.03	48.83	47.12
	Ino	43.94	44.23	43.72	41.95	48.65	49.52	48.42	46.69
	MOCHEG	45.62	46.53	45.40	43.84	52.01	53.26	51.77	50.04
	HGTMFC	**48.61**	**49.80**	**48.39**	**46.78**	**54.05**	**55.41**	**53.81**	**52.03**

Table 1: Overall fact checking result on MOCHEG. Bold indicates the best result, while underline denotes the second best.

By comparing the experimental outcomes, the multi-modal fact-checking framework HGTMFC that we develop has demonstrated superior performance in multi-modal version, text-only version and image-only version, particularly in terms of retrieval evidence, surpassing existing methodologies. This achievement highlights the effectiveness of our proposed model, marking a significant advancement in the current domain of multi-modal fact-checking.

Comparison to Multi-Modal Methods In the realm of multi-modal fact-checking, our HGTMFC method exhibits notable advantages over existing approaches such as Triple-Check. Specifically, in the experiments involving gold evidence, HGTMFC achieves an accuracy of 54.05% on the MOCHEG dataset, representing a substantial improvement of 2.04% over the MOCHEG method, which achieves 52.01%. Additionally, in experiments with retrieved evidence, HGTMFC attain an accuracy of 48.61%, surpassing the 45.62% accuracy of MOCHEG. There was also a marked improvement in F1 scores, underscoring HGTMFC's ability to more effectively integrate textual and visual information when processing multi-modal data, thereby enhancing the accuracy of fact-checking.

Several existing multi-modal fact-checking methods, such as Triple-Check and MOCHEG, determine the stance of each piece of evidence relative to the claim and then combine these stances. However, this simple fusion approach, while widely used, is insufficient to capture the higher-order relationships within and across different modalities due to the distinct semantic spaces in which different modalities exist and the absence of multi-modal fine-grained information.

Comparison to Text-Only Methods To further evaluate the performance of our model in text processing, we modify HGTMFC by removing the visual information processing module, resulting in HGTMFC-w text. Despite relying solely on textual evidence, this modified approach did not exhibit a significant drop in performance, achieving a LA score of 46.68% with retrieval evidence and 50.74% with gold evidence. This performance also surpasses that of other text-based fact-checking methods, such as GEAR, KGAT, and HESM, demonstrating that our approach remains competitive in text processing and maintains high accuracy even in the absence of visual information.

Model	Retrieval Evi		Gold Evi	
	LA	F1	LA	F1
Mean aggregation	46.06	44.18	50.66	48.63
Full self-attention	48.12	46.14	53.36	51.33
Homogeneous attention	46.19	44.18	51.02	49.02
w/o linegraph	48.16	46.35	53.48	51.51
HGTMFC	**48.61**	**46.78**	**54.05**	**52.03**

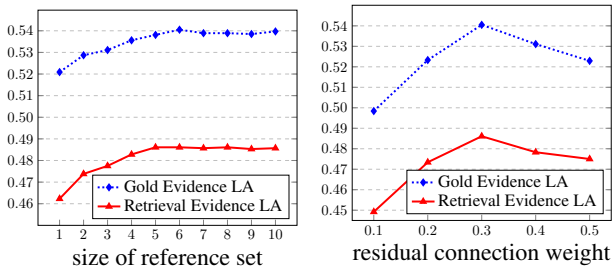
Table 2: Performance of ablation Study.

Comparison to Image-Only Methods Similarly, we remove textual evidence to focus exclusively on image processing, creating HGTMFC-w image. When compare with other image-based evidence inference models, HGTMFC-w image is superior to two other methods like Triple-Check-w image and MOCHEG-w image.

These results collectively demonstrate that HGTMFC significantly outperforms existing techniques in multi-modal, text-only, and image-only processing. Our method's ability to capture fine-grained evidence information enhances the accuracy of claim verification.

Ablation Study

In this subsection, we present an ablation study to assess the impact of various components on our model's performance. Table 2 shows the label accuracy and F1 score using gold evidence after removing components such as group attention, heterogeneous attention, and the line graph. Firstly, replacing intra-hyperedge aggregation based on group attention with average aggregation or full self-attention resulted in worse performance. Average aggregation, which averages



(a) Accuracy vs. p (b) Accuracy vs. β

Figure 4: Hyperparameter sensitivity analysis.

Model	Training Cost	Inference Cost	LA
Ino	32min	2min26s	48.65
MOCHEG	64min	2mins49s	52.01
HGTMFC	74min	3min17s	**54.05**

Table 3: Execution efficiency analysis of models

token embeddings, performed the worst, highlighting the importance of token interrelations. In contrast, group mixed attention provided a more robust hyperedge representation. Secondly, replacing heterogeneous attention with homogeneous attention caused a significant performance drop, indicating the importance of node and hyperedge diversity in integrating information from different feature spaces. Lastly, removing the line graph led to a decrease in performance, demonstrating its role in capturing low-order relationships between edges. These results confirm that each proposed component improves the framework's performance.

Hyperparameter Sensitivity Analysis

In this section, we investigate the sensitivity of HGTMFC to the core hyperparameters, the size of reference set p and the residual connection weight β. Figure 4(a) illustrates the performance curve of p. It can be observed from the figure that the performance of HGTMFC first increase and keep steady, as a larger number of reference set will capture more comprehensive semantics. From 4(b), the performance of HGTMFC initially increases and then decreases as we increase the weight β from 0.1 to 0.9. A larger β introduces more heterogeneous topological information into text token or image patch representations, thereby providing additional evidence for fact verification tasks. However, increasing β may also amplify noise from the introduction of other topological information. Therefore, optimizing this hyperparameter is crucial to maximize the model's performance.

Efficiency Analysis

We conduct an analysis of HGTMFC's execution efficacy, as shown in Table 3. Comparing it with two CLIP-based models, Ino and MOCHEG, we achieve the following observations. The HGTMFC model exhibits longer training and inference times compared to the Ino and MOCHEG. Both Ino and MOCHEG initially use the CLIP model to encode claims and evidence. In Ino, the embeddings are processed by MLP for classification, whereas in MOCHEG, the embeddings are directed to a stance detection module for classification. Consequently, the training duration is primarily influenced by the latter stages of their classifiers, with inference requiring the simultaneous use of both the CLIP model and the classification modules. In contrast, the HGTMFC model integrates CLIP within a hypergraph transformer, enabling end-to-end training. This integrated approach, while enhancing model performance, results in increased training and inference times relative to the other two models.

Related Work

Fact Verification

Fact verification assesses claim authenticity by retrieving evidence from verified text or image documents. The process involves three steps: evidence retrieval, claim verification, and explanation generation. Text-based models for claim verification like FEVER which proposed using Bi-LSTM to encode claims and evidence separately (Thorne et al. 2018) and BERT-based methods leveraging pre-trained models like BERT or ALBERT for robust text representations (Zhou et al. 2019; Subramanian and Lee 2020) are widely used. Additionally, graph neural networks have been applied in this context (Liu et al. 2019; Lu and Li 2020; Xu et al. 2022; Chen et al. 2022; Wu et al. 2023). Multi-modal approaches, such as using Consistency Check Network (Abdelnabi et al. 2022) for combining textual and visual data with large models (Du et al. 2023; Zhang et al. 2023b, 2025; Zhang, Zhang, and Pan 2022), have also been explored.

Hypergraph Learning

Hypergraph networks, which connect multiple nodes through a single hyperedge, are used to model complex relationships. This approach has been applied in clustering (Takai et al. 2020; Hu et al. 2021), classification (Sun et al. 2021; Wu et al. 2024), traffic prediction (Xu et al. 2024), representation learning (Jiang et al. 2024), and recommendation systems (Xia et al. 2021; Xia, Huang, and Zhang 2022; Tian et al. 2023). Hypergraph convolutional networks extend GCNs to process hypergraph-structured data (Zhang et al. 2022; Cai et al. 2022). In our task, hypergraphs help integrate higher-order information between claims and multi-modal evidence for improving fact verification.

Conclusion

In this study, we propose a novel approach for fine-grained multi-modal fact-checking by leveraging a Hypergraph Transformer to model high-order relationships between textual and visual evidence. Our framework integrates multi-scale features and utilizes a line graph to effectively capture complex interactions across different modalities. Experiments on benchmark datasets demonstrate that HGTMFC outperforms state-of-the-art methods, demonstrating its effectiveness in multi-modal evidence fusion and reasoning.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No. 62372057)

References

Abdelnabi, S.; Hasan, R.; Fritz, M.; and ". 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14940–14949.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Cai, D.; Song, M.; Sun, C.; Zhang, B.; Hong, S.; and Li, H. 2022. Hypergraph Structure Learning for Hypergraph Neural Networks. In *IJCAI*, 1923–1929.

Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; and Sun, X. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3438–3445.

Chen, Z.; Hui, S. C.; Zhuang, F.; Liao, L.; Li, F.; Jia, M.; and Li, J. 2022. Evidencenet: Evidence fusion network for fact verification. In *Proceedings of the ACM Web Conference 2022*, 2636–2645.

Dong, M.; and Kluger, Y. 2023. Towards understanding and reducing graph structural noise for GNNs. In *International Conference on Machine Learning*, 8202–8226. PMLR.

Du, W.-W.; Wu, H.-W.; Wang, W.-Y.; and Peng, W.-C. 2023. Team Triple-Check at Factify 2: Parameter-Efficient Large Foundation Models with Feature Representations for Multi-Modal Fact Verification. *arXiv preprint arXiv:2302.07740*.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hu, Y.; Li, X.; Wang, Y.; Wu, Y.; Zhao, Y.; Yan, C.; Yin, J.; and Gao, Y. 2021. Adaptive hypergraph auto-encoder for relational data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(3): 2231–2242.

Jiang, Y.; Gao, Y.; Zhu, Z.; Yan, C.; and Gao, Y. 2024. HyperRep: Hypergraph-Based Self-Supervised Multimodal Representation Learning.

Kim, J.; Park, S.; Kwon, Y.; Jo, Y.; Thorne, J.; and Choi, E. 2023. FactKG: Fact verification via reasoning on knowledge graphs. *arXiv preprint arXiv:2305.06590*.

Li, C.; Pang, B.; Liu, Y.; Sun, H.; Liu, Z.; Xie, X.; Yang, T.; Cui, Y.; Zhang, L.; and Zhang, Q. 2021. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 223–232.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, 163–179. Springer.

Li, C.; Wang, S.; Yu, P. S.; Zheng, L.; Zhang, X.; Li, Z.; and Liang, Y. 2018. Distribution distance minimization for unsupervised user identity linkage. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 447–456.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2019. Fine-grained fact verification with kernel graph attention network. *arXiv preprint arXiv:1910.09796*.

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Subramanian, S.; and Lee, K. 2020. Hierarchical evidence set modeling for automated fact extraction and verification. *arXiv preprint arXiv:2010.05111*.

Sun, X.; Yin, H.; Liu, B.; Chen, H.; Cao, J.; Shao, Y.; and Viet Hung, N. Q. 2021. Heterogeneous hypergraph embedding for graph classification. In *Proceedings of the 14th ACM international conference on web search and data mining*, 725–733.

Takai, Y.; Miyauchi, A.; Ikeda, M.; and Yoshida, Y. 2020. Hypergraph clustering based on pagerank. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1970–1978.

Tao, Z.; Wei, Y.; Wang, X.; He, X.; Huang, X.; and Chua, T.-S. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5): 102277.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355*.

Tian, Z.; Li, C.; Zuo, Z.; Wen, Z.; Sun, L.; Hu, X.; Zhang, W.; Huang, H.; Wang, S.; Deng, W.; et al. 2023. Pass: Personalized advertiser-aware sponsored search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4924–4936.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wu, H.; Li, N.; Zhang, J.; Chen, S.; Ng, M. K.; and Long, J. 2024. Collaborative contrastive learning for hypergraph node classification. *Pattern Recognition*, 146: 109995.

Wu, L.; Yu, D.; Liu, P.; Gao, C.; and Wang, Z. 2023. Heuristic heterogeneous graph reasoning networks for Fact Verification. *IEEE Transactions on Neural Networks and Learning Systems*.

Xia, L.; Huang, C.; and Zhang, C. 2022. Self-supervised hypergraph transformer for recommender systems. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2100–2109.

Xia, X.; Yin, H.; Yu, J.; Wang, Q.; Cui, L.; and Zhang, X. 2021. Self-supervised hypergraph convolutional networks for session-based recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4503–4511.

Xu, C.; Wei, Y.; Tang, B.; Yin, S.; Zhang, Y.; Chen, S.; and Wang, Y. 2024. Dynamic-group-aware networks for multi-agent trajectory prediction with relational reasoning. *Neural Networks*, 170: 564–577.

Xu, W.; Wu, J.; Liu, Q.; Wu, S.; and Wang, L. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, 2501–2510.

Yao, B. M.; Shah, A.; Sun, L.; Cho, J.-H.; and Huang, L. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743.

Yin, Y.; Meng, F.; Su, J.; Zhou, C.; Yang, Z.; Zhou, J.; and Luo, J. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. *arXiv preprint arXiv:2007.08742*.

Zhang, L.; Zhang, X.; Li, C.; Zhou, Z.; Liu, J.; Huang, F.; and Zhang, X. 2024a. Mitigating Social Hazards: Early Detection of Fake News via Diffusion-Guided Propagation Path Generation. In *ACM Multimedia 2024*.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical cross-modality semantic correlation learning model for multi-modal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11676–11684.

Zhang, L.; Zhang, X.; Zhou, Z.; Huang, F.; and Li, C. 2024b. Reinforced adaptive knowledge learning for multi-modal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16777–16785.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Wang, S.; Philip, S. Y.; and Li, C. 2024c. Early Detection of Multimodal Fake News via Reinforced Propagation Path Generation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Yu, P. S.; and Li, C. 2025. Knowledge-aware multimodal pre-training for fake news detection. *Information Fusion*, 114: 102715.

Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J. B.; Zhang, Y.; Xie, X.; Wang, H.; and Kim, S. 2023a. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 168–176.

Zhang, Y.; Tao, Z.; Wang, X.; and Wang, T. 2023b. Ino at factify 2: Structure coherence based multi-modal fact verification. *arXiv preprint arXiv:2303.01510*.

Zhang, Z.; Feng, Y.; Ying, S.; and Gao, Y. 2022. Deep hypergraph structure learning. *arXiv preprint arXiv:2208.12547*.

Zhao, J.; Li, C.; Wen, Q.; Wang, Y.; Liu, Y.; Sun, H.; Xie, X.; and Ye, Y. 2021. Gophormer: Ego-graph transformer for node classification. *arXiv preprint arXiv:2110.13094*.

Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.