Instruction-Tuned LLMs Meet Cross-Modal Label Propagation: A Cross-Modal Framework for Fake News Detection

Anonymous ACL submission

Abstract

The proliferation of multimodal fake news severely undermines the information ecosystem, and its accurate detection has become a core research topic in natural language processing and multimedia analysis. Existing approaches integrating LLM-based pseudo-label generation and label propagation have shown promise but suffer two limitations: first, LLMs lack task adaptability due to the absence of task-specific fine-tuning; second, pseudo-labels generated by LLMs solely from text data result in notable modal bias against the multimodal features relied on by the detection task. To address these issues, we first conduct task-specific multimodal collaborative instruction fine-tuning on LLMs, which addresses the modal bias at its root and enhances pseudo-label quality. We then design a multimodal feature transformation alignment module to tackle the secondary modal mismatch between general multimodal features and pseudo-labels generated by fine-tuned LLMs. This work presents a multimodal LLM fine-tuning paradigm, a cross-modal label propagation mechanism integrating the feature alignment module, node labeling rules, and a pseudo-label confidence-based linear weighting strategy, and the LLM-Tuned Cross-Modal Label Propagation Framework (LLM-T-CMLP). Experiments on three public benchmark datasets demonstrate that our framework outperforms current state-of-the-art (SOTA) baselines by a notable margin, fully confirming the effectiveness of our proposed methods.

1 Introduction

The rapid spread of multimodal fake news misleads public perception and erodes the foundation of social trust, making its efficient identification an urgent need in natural language processing and multimedia analysis (Wu et al., 2021; Chen et al., 2022; Zhang et al., 2023; Wei and Wu, 2024; Cao et al., 2025). In recent years, Large Language Models (LLMs), by virtue of their powerful semantic understanding and reasoning capabilities (Greco et al., 2025; Rashkin et al., 2025), have provided new technical support for fake news detection (Hu et al., 2025; Zhang et al., 2025a; Tong et al., 2025; Hu et al., 2024); meanwhile, label propagation technology has demonstrated unique advantages in tapping into the value of unlabeled data and optimizing classification performance (Zhu and Ghahramani, 2002; Iscen et al., 2019). The integration of the two has paved a new path for multimodal fake news detection. As shown in Figure 1(a), a study (Hu et al., 2025) innovatively integrated LLM pseudo-label generation with label propagation and applied it to multimodal fake news detection, achieving good results. However, the framework has two core limitations: first, it does not conduct task-specific image-text collaborative instruction fine-tuning for LLM, leaving the model lacking adaptability to multimodal fake news detection; second, pseudo-labels are generated solely based on the single text modality, leading to modal bias with the image-text fused features output by the encoder and resulting in semantic disconnection between labels and features. To address these two core issues, we first introduce task-specific image-text collaborative instruction fine-tuning for LLM to enhance the task

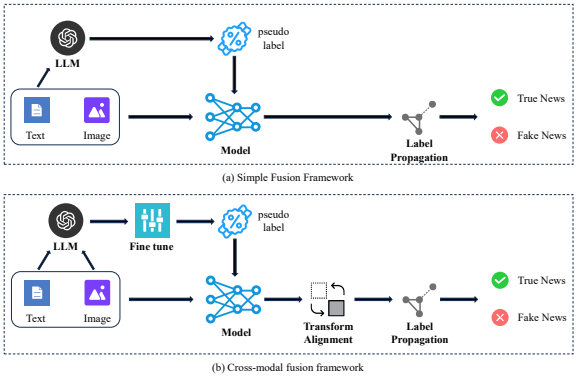

Figure 1: Description of different fusion framework.

adaptability of pseudo-labels. Furthermore, we found that the multimodal features output by the encoder are general-purpose, forming a secondary modal mismatch with the pseudo-labels that have strong task specificity after fine-tuning. To this end, we additionally designed a multimodal feature transformation alignment module to completely resolve the modal mismatch problem. Based on this, our core contributions are as follows:

- We propose a multimodal LLM instruction fine-tuning paradigm. Through customized image-text collaborative instructions and cross-modal training, LLM judges the authenticity of news by combining images and text, fundamentally solving the modal bias of pseudo-labels relying solely on text and ignoring images, and generating high-precision pseudo-labels.
- We design a cross-modal label propagation mechanism: the multimodal feature transformation alignment module maps multimodal features to a unified task-adapted semantic space to bridge modal mismatch; we also design a node labeling rule and a linear weighting strategy based on pseudo-label confidence to mitigate errors caused by fluctuations in pseudo-label quality, thereby improving the accuracy and stability of label propagation.
- We propose the LLM-Tuned Cross-Modal Label Propagation Framework (LLM-T-CMLP). Experiments on three public benchmark datasets show that this framework significantly outperforms current state-of-the-art (SOTA) baseline models, fully validating the effectiveness of the proposed methods.

2 Related work

This section provides an overview of related work on multimodal fake news detection, large language models, and label propagation.

2.1 Multimodal Fake News Detection

In the early stage of fake news detection, textual unimodal analysis was the core (Wang et al., 2018a). With the explosion of multimedia content on social media, research has shifted to multimodal fusion (Wu et al., 2021; Suryavardan et al., 2023; Li et al., 2025), whose core lies in integrating complementary information from text and images to improve performance. Existing methods are mainly categorized into three types: modeling modal dependencies via attention mechanisms (Qian et al., 2021; Zhang et al., 2025b), constructing multimodal graphs with graph neural networks (GNNs) to capture sample correlations (Zhao et al., 2023; Wang et al., 2020), and learning cross-modal unified representations with generative models (Khattar et al., 2019).

In recent years, several studies (Hu et al., 2025; Tong et al., 2025) have attempted to incorporate Large Language Models (LLMs) to enhance detection performance. However, these methods rely solely on textual semantics and mostly adopt native calls of general-purpose APIs, lacking adaptation and optimization for the characteristics of multimodal detection tasks. This results in a disconnect between LLM outputs and task requirements, limiting performance improvements.

2.2 Large Language Models

With the growing demand for multimodal tasks, multimodal Large Language Models (LLMs) that integrate textual and visual understanding have emerged as a research hotspot, effectively breaking through the unimodal limitations of traditional LLMs (Zang et al., 2025; Yang et al., 2023).

Closed-source models, represented by GPT-4o (Hurst et al., 2024), achieve deep image-text collaborative reasoning through end-to-end training, delivering outstanding performance on multimodal tasks. However, their "black-box" nature, access restrictions, and invocation costs have constrained customized research in the academic community.

In the open-source domain, Qwen2.5-VL (Bai et al., 2025) stands out: it has superior performance and supports full-process fine-tuning. Besides, the 7B parameter scale significantly lowers the computational barrier for academic research. Therefore, we attempt to leverage Qwen2.5-VL-7B to enhance fake news detection performance.

2.3 Label Propagation

Label Propagation (LP), a classic method in semi-supervised learning, achieves the classification of unlabeled samples by propagating label information across graphs (Zhu and Ghahramani, 2002). Recent works have integrated label propagation with graph neural networks (GNNs) (Kipf, 2016), demonstrating promising performance in various application scenarios (Duan et al., 2024; Liu et al., 2024; Choi et al., 2025), particularly in the field of multimodal fake news detection where encour-

aging results have been achieved (Hu et al., 2025; Zhao et al., 2023).

3 Method

We propose the LLM-Tuned Cross-Modal Label Propagation Framework (LLM-T-CMLP), illustrated in Figure 2. It comprises two core components: LLM-based pseudo-label generation and cross-modal label propagation. First, we design prompts for the image-text data and instruction-tune the base LLM to obtain LLM-T. LLM-T then generates pseudo-labels and confidence scores for unlabeled samples. Next, CLIP encodes image-text data into multimodal features, which are aligned cross-modally and fused with pseudo-labels and confidence scores. Finally, graph neural networks (GNNs) complete label propagation to yield multimodal fake news classification results.

3.1 LLM Fine-Tuning for Pseudo-Label Generation

To improve the accuracy of LLM-generated pseudo-labels and their adaptability to multimodal features, we propose two optimization strategies: (1) Domain-adaptive instruction tuning: We perform task-specific instruction tuning on Qwen2.5-VL-7B using multimodal fake news data (image-text pairs with authenticity labels) to improve pseudo-label accuracy and reliability. (2) Adaptive instructional prompt design: We design prompts tailored for multimodal inputs, guiding the model to fully fuse visual-textual information for unified, concise, and accurate pseudo-labels and provide the judgment criteria.

Specifics on data preprocessing, standardized prompt format, and training strategies are in Appendix A, B, and C.

3.2 Multimodal Feature Extraction and Cross-Modal Graph Construction

For fake news detection that requires joint utilization of textual semantics and visual information, we first adopt CLIP (Radford et al., 2021) for multimodal feature extraction.

For each image-text news sample, CLIP's dual-branch encoder generates modality-specific high-dimensional features: The visual branch encodes news images into image feature vectors $v_i \in \mathbb{R}^{d_v}$, capturing color, texture, and high-level semantic information. The textual branch encodes news body text into textual feature vectors $t_i \in \mathbb{R}^{d_t}$, capturing semantic logic, emotional tendencies, and key entity information.

We concatenate these two feature vectors to form a unified multimodal representation $x_i \in \mathbb{R}^{d_t+d_v}$, whose mathematical expression is:

$$x_i = t_i \oplus v_i \tag{1}$$

Where $\oplus$ represents the feature dimension concatenation operation.We then construct a cross-modal association graph following the paradigm of FCN-LP (Zhao et al., 2023), where nodes represent news samples and edges denote strong associations. Nodes directly adopt the unified multimodal features x_i extracted above. Edges are determined by three similarity metrics with a threshold $\theta = 0.95$: *1) Comprehensive similarity*: Cosine similarity between multimodal features x_i and x_j, reflecting the overall text-vision association. *2) Cross-modal similarity*: Bidirectional matching between $t_i\&v_j$ and $v_i\&t_j$, capturing cross-modal semantic correlations. *3) Intra-modal similarity*: Similarities between image features $v_i\&v_j$ and textual features $t_i\&t_j$, focusing on single-modality direct associations.

An undirected edge is added if any similarity exceeds θ, retaining only strong connections. This graph comprehensively covers multi-dimensional associations and ensures valid label propagation.

3.3 Transformation Alignment of Node Multimodal Features

Multimodal features extracted by CLIP have achieved general semantic alignment between textual and image features. However, to eliminate modal bias with pseudo-labels generated for fake news detection tasks, a task-specific refined alignment module still needs to be designed. Specifically, for the original multimodal feature x_i, we map the general features of its textual feature vector t_i and image feature vector v_i to a task-relevant discriminative space via lightweight MLPs, respectively:

$$t_i' = \text{LayerNorm}(\sigma(W_t t_i)) \tag{2}$$

$$v_i' = \text{LayerNorm}(\sigma(W_v v_i)) \tag{3}$$

where $W_t \in \mathbb{R}^{d_t \times (d_t+d_v)}$ and $W_v \in \mathbb{R}^{d_v \times (d_t+d_v)}$ denotes the task-adaptive linear transformation weight. σ is an activation function that enhances the non-linear discriminative capability of features. LayerNorm normalizes the multimodal feature distributions to similar ranges, avoiding information being overwhelmed due to scale differences.

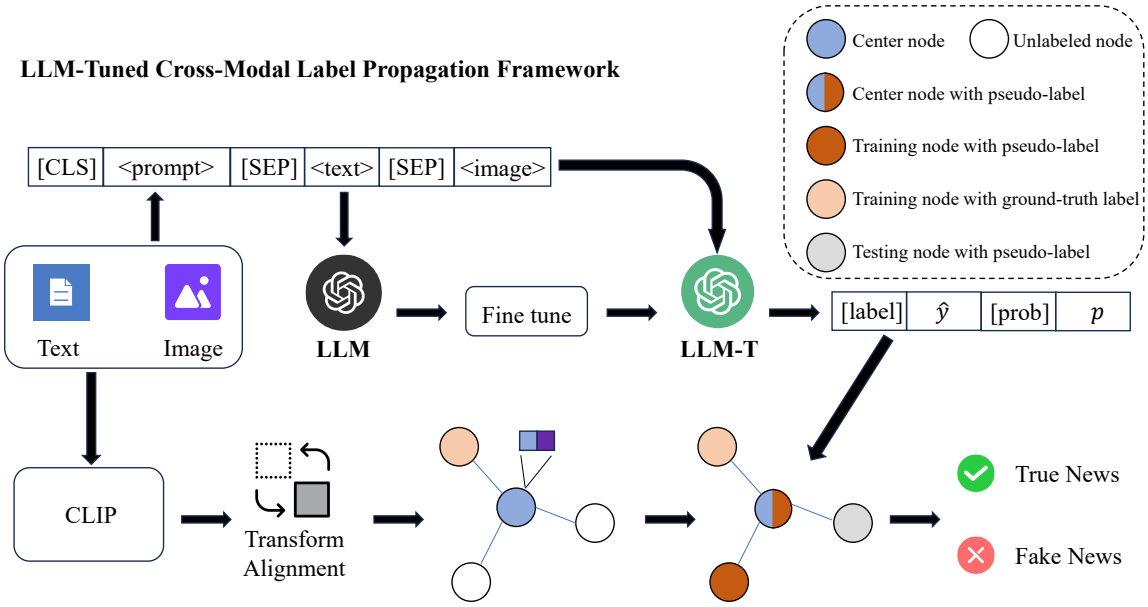

Figure 2: Overall framework of the LLM-T-CMLP.

Subsequently, we introduce learnable modal weights $w = [w_v, w_t]^T$ (with the initial value set to $w = [1, 1]^T$) to fuse the aligned image and textual feature vectors:

$$\hat{w} = \text{softmax}(w) \tag{4}$$

$$h_i = (\hat{w}_v v_i' + \hat{w}_t t_i') \oplus x_i \tag{5}$$

where $\oplus$ represents residual connection (He et al., 2016).

3.4 Node Label Assignment Rules

Effective pseudo-label utilization is critical for semi-supervised node classification. We propose refined node label assignment rules to support subsequent fusion of pseudo-labels and node features.

We use the fine-tuned LLM to generate predicted pseudo-labels $\hat{y}$ and confidence scores $p \in [0, 1]$, where p quantifies the LLM's prediction confidence. To distinguish label sources (ground-truth/pseudo-labels) and clarify their training roles, we construct a full-node label tensor $Y_{\text{all}} \in \{0, 1\}^N$ (N denotes total nodes) and a confidence tensor $P_{\text{all}} \in [0, 1]^N$. Specific assignment rules are as follows: *1) Ground-truth training nodes*: Random subset of the training set, assigned ground-truth labels $Y_{\text{all}}[i] = y_i^{\text{train}} \in \{0, 1\}$ with $P_{\text{all}}[i] = 1.0$ (providing noise-free supervision signals). *2) Pseudo-training nodes*: Remaining training set nodes, with ground-truth labels hidden during training. Assigned pre-generated pseudo-labels $Y_{\text{all}}[i] = \hat{y}_i^{\text{train}} \in \{0, 1\}$ and confidence $P_{\text{all}}[i] = p_i^{\text{train}}$. Ground-truth labels are used for loss calculation, while pseudo-labels enable feature enhancement to simulate weak supervision. *3) Pseudo-test nodes*: Test set nodes, assigned pseudo-labels $Y_{\text{all}}[i] = \hat{y}_i^{\text{test}} \in \{0, 1\}$ and confidence $P_{\text{all}}[i] = p_i^{\text{test}}$. These expand training samples, introduce latent supervision from the test set, and alleviate train-test distribution shift.

It is important to note that the pseudo-labels of test set nodes are independently generated by the LLM fine-tuned on the training set, without accessing any true labels of the test set during the generation process, thus ensuring no label leakage.

3.5 Confidence-Weighted Pseudo-Label Fusion Mechanism

Our method adopts a confidence-weighted pseudo-label fusion mechanism, which deeply couples pseudo-label information with node features and graph structure to achieve the expansion and enhancement of supervision signals.

3.5.1 Confidence-Weighted Label Embedding Generation

To balance the impact of labels with different confidence levels on label propagation, we design a label embedding module to convert Y_{all} and P_{all} into label embedding vectors $E \in \mathbb{R}^{N \times d}$ (where d denotes the embedding dimension, consistent with the node feature dimension). For any node i, weighted embedding is performed based on the

label confidence:

$$E[i] = P_{\text{all}}[i] \cdot e \tag{6}$$

where $e \in \mathbb{R}^d$ is a learnable label embedding vector. This formula dynamically balances the embedding contribution of the label through confidence level.

Furthermore, we adjust the sign of the embedding vector by combining label values to enhance the consistency between labels and features:

$$E'[i] = E[i] \cdot (2 \cdot Y_{\text{all}}[i] - 1) \tag{7}$$

where $(2 \cdot Y_{\text{all}}[i] - 1)$ is the label sign coefficient. With $Y_{\text{all}}[i] \in \{0, 1\}$, this coefficient maps label 0 to -1 and label 1 to 1, ensuring alignment between the embedding direction and label semantics.

3.5.2 Residual Fusion of Label Embeddings and Node Features

We perform residual fusion between the adjusted label embeddings $E'[i]$ and the node features h_i aligned via multimodal transformation:

$$h'_i = h_i \oplus E'[i] \tag{8}$$

This operation directly injects label information into node features, enabling the subsequent graph propagation process to both leverage the structural information of original features and carry the semantic signals of labels. For nodes with high-confidence pseudo-labels, their label embeddings have higher weights, allowing them to transmit more prominent supervision signals to neighbors during propagation; for low-confidence nodes, the contribution of their embeddings is weakened, preventing noise diffusion.

Node features fused with pseudo-label information are fed into stacked TransformerConv (Shi et al., 2020) propagation networks. By virtue of neighborhood correlations in the graph structure, these networks achieve efficient transmission and aggregation of label signals, providing structured feature support for the final node classification. Finally, we adopt a cross-entropy loss function to optimize the model output.

4 Experiment

4.1 Datasets

Twitter (Boididou et al., 2015): This dataset is derived from the MediaEval Verifying Multimedia Use benchmark dataset (Boididou et al., 2015), designed for social media fake content detection tasks. It contains approximately 17,000 unique tweets associated with multiple types of events. Each tweet includes textual content and relevant images, with annotations indicating content authenticity. Following the setup in FCN-LP (Zhao et al., 2023), we split the dataset into a training set and a test set, consisting of 15,000 and 2,000 tweets, respectively.

PHEME (Zubiaga et al., 2017): Composed of tweets related to breaking news on the Twitter platform, this dataset focuses on 5 major breaking news events. Each event includes a large number of text-image sample pairs, all annotated for content authenticity. Following the setup in FCN-LP (Zhao et al., 2023), we use 1,414 and 608 tweets as the training set and test set, respectively.

Weibo (Jin et al., 2017): Collected from Sina Weibo, a Chinese social media platform, this dataset is split following the setup in FCN-LP (Zhao et al., 2023), with 4,141 tweets for the training set and 1,125 tweets for the test set.

It's worth noting that the text in both the Twitter and PHEME datasets is in English, while the text in the images in the Weibo dataset is in English, and the text in the images is in Chinese.

4.2 Baseline Models

To comprehensively verify the performance advantages of the proposed LLM-T-CMLP framework, this paper selects various state-of-the-art (SOTA) methods as baselines, which are specified as follows:

1) Traditional multimodal feature fusion methods: LLM-independent, integrating text-image information through handcrafted feature extractors and fusion modules. These include EANN (Wang et al., 2018b), SpotFake (Singhal et al., 2019), MVAE (Khattar et al., 2019), SAFE (Zhou et al., 2020), MCAN (Wu et al., 2021), HMCAN (Qian et al., 2021), FCN (Zhao et al., 2023), and FCN-LP (Zhao et al., 2023).

2) LLM-based prediction methods: Prompt-driven, directly generating labels via LLMs. These include GPT-4o (Hu et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025) (without fine-tuning), and Qwen2.5-VL-7B-tuned (Bai et al., 2025) (fine-tuned on the dataset).

3) LLM-enhanced label propagation methods: To ensure fairness, we uniformly adopt the Qwen2.5-VL series as the LLM backbone. We replace GPT-4o in the original GLPN-LLM (Hu et al., 2025) with Qwen2.5-VL-7B-tuned while retaining its original mechanism, resulting in GLPN-

Model	Twitter				PHEME				Weibo			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
EANN	71.53±0.91	71.38±1.23	63.82±2.11	68.91±1.58	70.17±0.79	71.28±1.32	67.36±2.17	69.10±1.83	79.18±0.76	80.31±1.23	78.52±0.32	79.44±2.13
SpotFake	77.16±1.57	75.32±1.14	87.83±0.63	85.14±0.07	81.37±2.38	79.53±2.27	81.22±2.43	79.43±0.75	86.39±2.51	86.12±0.53	87.17±2.63	83.22±1.41
MVAE	74.56±1.58	80.15±2.69	76.34±0.83	81.57±1.98	77.83±1.27	73.82±2.05	73.45±2.62	72.21±0.54	71.86±0.25	70.32±0.69	70.32±2.84	70.53±1.60
SAFE	76.66±3.00	76.32±1.94	75.41±2.12	76.37±2.85	81.25±1.34	79.22±2.76	79.11±1.45	79.69±2.67	84.91±2.12	83.81±1.58	82.19±1.16	83.01±1.70
MCAN	80.91±2.33	82.68±2.48	76.67±0.94	82.26±1.32	80.74±1.89	79.21±2.23	79.64±1.53	80.15±0.86	86.50±3.00	88.10±2.10	84.60±1.80	86.15±1.60
HMCAN	83.91±1.49	81.68±2.08	84.67±1.21	82.57±1.62	86.36±1.83	83.18±1.41	83.81±2.51	83.49±1.07	86.75±2.95	88.40±3.00	84.65±1.80	87.20±1.20
FCN	82.86±1.27	78.64±1.68	87.39±0.85	82.78±0.47	80.36±1.93	84.43±1.27	89.12±0.12	86.71±1.88	82.92±0.54	83.17±1.00	88.45±2.13	86.74±0.41
FCN-LP	85.32±2.56	81.52±2.82	89.32±0.99	85.24±1.93	84.68±0.81	86.32±1.55	89.85±1.22	87.97±0.88	84.47±1.66	88.41±0.26	91.18±0.69	89.78±0.84
GPT-4o	75.39±3.32	75.66±3.44	80.92±2.83	78.20±5.66	74.38±6.68	78.66±5.31	75.16±4.32	76.87±5.65	80.86±3.11	82.16±2.86	81.33±3.66	81.75±2.95
Qwen2.5-VL-7B	54.14±0.87	58.15±0.65	70.25±0.73	63.63±0.66	55.28±1.23	69.64±0.36	66.05±2.23	67.78±1.44	76.93±1.03	82.73±0.36	88.36±1.12	85.45±0.71
Qwen2.5-VL-7B-tuned	79.62±0.09	79.40±0.08	86.84±0.34	82.96±0.11	84.10±1.46	87.50±1.25	90.66±0.75	89.05±1.00	87.84±0.65	94.55±0.32	89.29±0.60	91.85±0.45
GLPN-LLM-T	86.14±0.43	78.44±0.63	**99.63±0.32**	87.77±0.26	84.93±0.41	89.93±0.56	88.81±0.85	89.36±0.32	86.65±0.32	90.79±0.20	91.92±0.69	91.35±0.26
LLM-T-CMLP	**90.71±0.44**	**94.19±1.46**	90.60±1.17	**92.36±1.28**	**86.69±0.08**	**90.23±0.16**	**91.20±0.33**	**90.72±0.08**	**91.10±0.14**	**95.83±0.26**	**92.42±0.47**	**94.09±0.12**

Table 1: Performance comparison on the Twitter, PHEME, and Weibo datasets. The highest value in each column is marked in bold.

Method	Twitter				PHEME				Weibo			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
LLM	54.14±0.87	58.15±0.65	70.25±0.73	63.63±0.66	55.28±1.23	69.64±0.36	66.05±2.23	67.78±1.44	76.93±1.03	82.73±0.36	88.36±1.12	85.45±0.71
LLM-CMLP	81.15±0.82	79.08±2.17	87.33±2.82	84.70±0.41	84.21±0.21	87.71±1.23	90.59±1.35	89.11±0.02	85.56±0.63	91.12±0.94	89.95±0.92	90.52±0.40
LLM-T	79.62±0.09	79.40±0.08	86.84±0.34	82.96±0.11	84.10±1.46	87.50±1.25	90.66±0.75	89.05±1.00	87.84±0.65	94.55±0.32	89.29±0.60	91.85±0.45
LLM-T-CMLP (- -)	83.29±4.05	85.57±6.32	86.19±8.47	85.47±3.72	84.44±0.66	88.14±1.36	90.39±1.72	89.23±0.49	86.45±0.30	90.82±0.73	91.61±1.24	91.20±0.26
LLM-T-CMLP (-)	88.33±0.36	90.04±0.37	89.46±0.39	89.75±0.30	86.03±0.09	90.08±0.40	90.36±0.42	90.22±0.05	90.36±0.04	95.79±0.05	91.46±0.05	93.57±0.03
LLM-T-CMLP	**90.71±0.44**	**94.19±1.46**	**90.60±1.17**	**92.36±1.28**	**86.69±0.08**	**90.23±0.16**	**91.20±0.33**	**90.72±0.08**	**91.10±0.14**	**95.83±0.26**	**92.42±0.47**	**94.09±0.12**

Table 2: Ablation study. The highest value in each column is marked in bold.

LLM-T. Meanwhile, we construct LLM-CMLP (Qwen2.5-VL-7B) and LLM-T-CMLP (Qwen2.5-VL-7B-tuned) to highlight the advantages of the CMLP mechanism and the value of fine-tuning through two sets of comparisons, respectively.

Detailed introductions to the baseline models are provided in Appendix D.

4.3 Implementation Details

For consistent comparison with baselines, we follow the settings of the prior work FCN-LP (Zhao et al., 2023) when constructing the cross-modal tweet graph. We adopt AdamW as the optimizer with a learning rate of 5e-3 and a label rate of 0.4, training the model for a total of 500 epochs. The model is trained five independent times, and we report the mean values and standard deviations of accuracy, precision, recall, and F1-score.

4.4 Performance Comparison

Performance results are presented in Table 1. Based on the experimental results, we draw the following conclusions:

- CMLP (cross-modal label propagation) delivers significant incremental value: LLM-T-CMLP outperforms GLPN-LLM-T on Qwen2.5-VL-7B-tuned, improving fake news detection accuracy and generalization via precise cross-modal consistency capture.

- Task-specific fine-tuning is critical for multimodal LLMs: Fine-tuned models (Qwen2.5-VL-7B-tuned, LLM-T-CMLP) outperform non-fine-tuned counterparts across metrics, enhancing domain adaptation and key feature recognition.

- LLM-T + cross-modal label propagation outperforms traditional approaches: LLM-T-CMLP surpasses traditional multimodal fusion and pure LLM-T methods, validating the hybrid paradigm's advancement for multimodal fake news detection.

4.5 Ablation Study

To further validate the effectiveness of each component in LLM-T-CMLP, we perform ablation studies on three benchmark datasets (Twitter, PHEME, Weibo), with results in Table 2. Model definitions: LLM-CMLP: Uses pseudo-labels from the original LLM. LLM-T-CMLP: Full framework. LLM-T-CMLP (- -): Ablation variant w/o multimodal transformation alignment, label assignment rules & confidence weighting strategy, using the label fusion method of GLPN-LLM (Hu et al., 2025). LLM-T-CMLP (-): Ablation variant w/o multimodal transformation alignment.

The core experimental conclusions are as follows: *1) Task-specific LLM fine-tuning yields significant value*: LLM-T (fine-tuned) delivers stable metric gains across datasets vs. non-fine-tuned LLMs—e.g., 47% higher accuracy and 30% higher F1-score on Twitter—confirming adaptability to

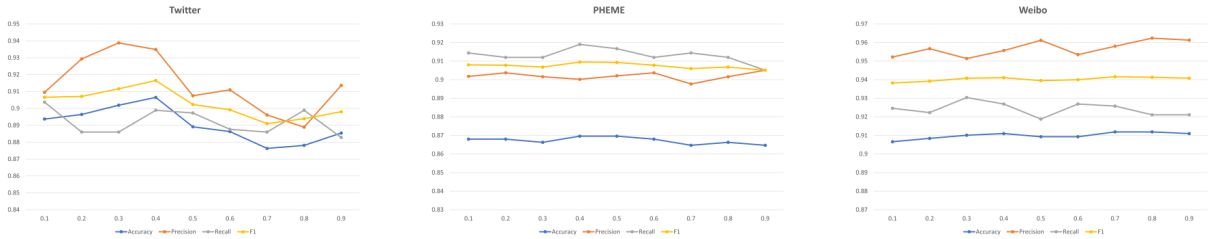

Figure 3: The effect of the label rate on the Twitter, PHEME, and Weibo datasets.

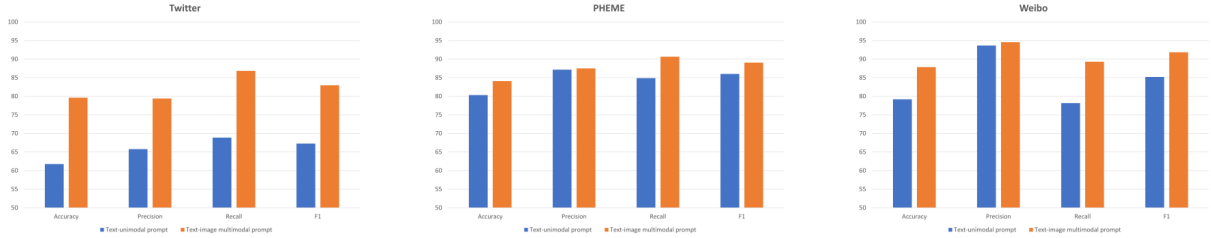

Figure 4: Impact of different prompt designs on performance.

fake news detection and enhanced multimodal feature capture. *2) The CMLP mechanism shows clear advantages*: Integrated with LLM-T, it boosts all metrics notably, directly validating its superiority. *3) Pseudo-label quality is critical*: LLM-T-CMLP outperforms LLM-CMLP, highlighting pseudo-label quality's major impact on final performance and revealing synergy between multimodal LLM fine-tuning and CMLP. *4) All core components are effective*: LLM-T-CMLP (- -) exhibits inferior performance to LLM-T-CMLP (-), and LLM-T-CMLP (-) in turn underperforms the full framework, confirming that these components are both indispensable and effective.

4.6 Hyperparameter Analysis

We investigate the impact of the proportion of nodes randomly selected from the training set to use ground-truth labels (label rate) on model performance, as shown in Figure 3. Experiments demonstrate that the model achieves optimal performance when the label rate is set to 0.4. This result stems from the comprehensive influence of the label rate on the quality of supervision signals, pseudo-label gains, and graph propagation efficiency: a label rate of 0.4 not only provides stable anchor points for classification boundaries through 40% of ground-truth training nodes but also reserves sufficient learning space for pseudo-labels. This allows pseudo-labels to effectively expand supervision coverage and capture potential associations between nodes.

4.7 Prompt Design Analysis

To investigate the impact of multimodal Prompt design on fake news detection tasks, we construct two types of comparative Prompts: *1) Text-unimodal prompt*: Only guides the model to judge authenticity based on news textual content, without any cross-modal collaborative guidance logic, and does not rely on image information throughout the process. *2) Text-image multimodal prompt*: Proactively guides the model to fuse textual semantics and image visual information for authenticity evaluation through explicit cross-modal collaborative evaluation instructions.

The performance comparison results of the two types of Prompts are shown in Figure 4. Experiments demonstrate that the Text-image multimodal prompt outperforms the Text-unimodal prompt significantly across the Twitter, PHEME, and Weibo datasets. This advantage stems from the Text-image multimodal prompt's ability to fully leverage the complementary information of text and images—for example, accurately identifying multimodal fake news features that are difficult to detect with unimodal methods, such as "authentic text but tampered images" and "false text but misleading images." In contrast, the Text-unimodal prompt only relies on textual information, failing to cover such key discriminative clues and resulting in limited performance.

Notably, the performance gap between the two types of Prompts is the most pronounced on the Twitter dataset. We speculate this is related to

the dataset's sample characteristics: compared to PHEME and Weibo, Twitter contains a higher proportion of samples that require text-image collaboration for accurate judgment. For such samples, the Text-unimodal prompt is prone to misjudgment due to lack of information, while the cross-modal fusion advantage of the Text-image multimodal prompt is maximized. Detailed designs of the specific prompts are provided in Appendix B.

4.8 Visualization

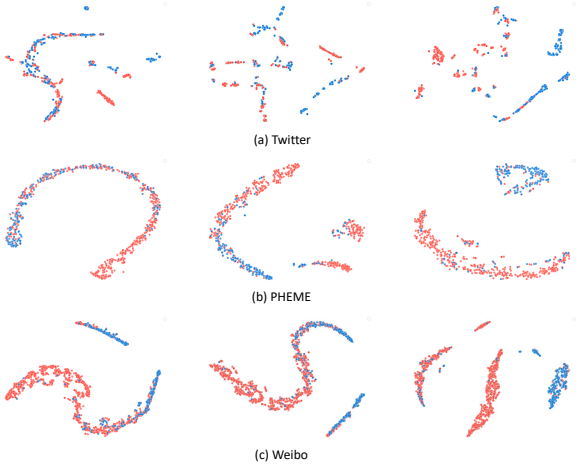

Figure 5: t-SNE visualizations of feature embeddings on the test set.

Figure 5 presents the t-SNE visualization results of feature embeddings from three models on the test set, with the three columns corresponding to the embedding effects of GLPN-LLM-T, LLM-CMLP, and LLM-T-CMLP, respectively. In terms of the separation degree between fake and real news, the clusters of fake and real news in GLPN-LLM-T and LLM-CMLP show significant overlap, while LLM-T-CMLP achieves a notably improved separation effect. Specifically, on the PHEME dataset, the discriminability between fake and real news by LLM-T-CMLP is particularly prominent. This result indicates that both the instruction tuning of LLM and the proposed cross-modal label propagation framework CMLP can exert a significant effect on enhancing the feature distinguishability between fake and real news.

4.9 Case Study

As illustrated in Figure 6, we visually demonstrate the classification performance of four models (LLM, LLM-T, GLPN-LLM-T, and LLM-T-CMLP) through typical cases: The LLM yields an incorrect classification result with a confidence score as high as 0.9, exhibiting strong overconfidence in its misjudgment. After instruction tuning, LLM-T enhances its discriminative ability for domain data and successfully outputs the correct classification result, but with a confidence score of only 0.3, indicating the model lacks sufficient confidence in the correct outcome. GLPN-LLM-T adopts pseudo-labels generated by LLM-T for label propagation. Due to the absence of the multimodal transformation and alignment mechanism as well as the confidence-weighted mechanism, it suffers from neighbor interference and produces incorrect classifications. In contrast, our proposed LLM-T-CMLP model not only achieves correct classification but also maintains a high confidence score, delivering superior overall performance.

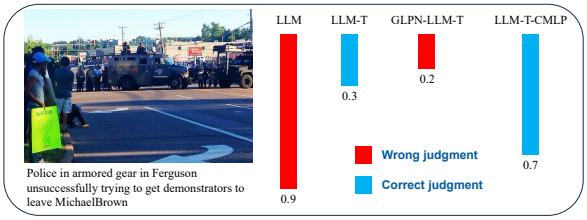

Figure 6: Case study of fake news detection.

5 Conclusion

This paper proposes an LLM-Tuned Cross-Modal Label Propagation Framework (LLM-T-CMLP) for multimodal fake news detection tasks. Our method performs instruction tuning on domain datasets using multimodal LLMs to generate high-quality pseudo-labels. We also design the CMLP framework, which effectively enhances the integration capability with pseudo-labels through multimodal transformation and alignment, as well as a confidence-weighted strategy. The effective combination of LLM-T and CMLP fully explores the associations in multimodal data, significantly improving the performance of fake news detection. In future work, we plan to expand multimodal inputs by incorporating richer modalities. Another future direction is to fuse external knowledge (e.g., knowledge graphs) to enhance LLMs' discriminative ability for fake information.

Limitations

Dataset Limitation: Existing datasets are predominantly centered on Chinese and English social media platforms, lacking fake news samples covering multiple languages, regions (e.g., non-English-

speaking countries), and domains (e.g., technology, healthcare). Thus, their generalization ability remains to be verified.

Model Efficiency: While the Qwen2.5-VL model with 7B parameters significantly reduces computational cost compared to GPT-4o, its complexity remains non-negligible in resource-constrained deployment environments. Future research can explore models with fewer parameters or model distillation techniques to optimize efficiency while maintaining performance.

Ethical Considerations and Risks

Beyond technical limitations, our work entails potential societal risks. Firstly, the model's performance may degrade on underrepresented languages or cultural contexts, raising fairness concerns. We explicitly acknowledge this bias and encourage future work to incorporate more diverse data. Secondly, while we employ efficient fine-tuning to reduce computational cost, the environmental impact of large-scale model deployment remains a consideration. Lastly, the underlying technology could be repurposed to generate sophisticated multimodal misinformation. We emphasize that our research is solely aimed at detection and support the development of ethical guidelines against malicious use.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc Tien Dang Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, and 1 others. 2015. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015*, volume 1436. CEUR-WS.

Biwei Cao, Qihang Wu, Jiuxin Cao, Bo Liu, and Jie Gui. 2025. External reliable information-enhanced multimodal contrastive learning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 31–39.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

Jinhyeok Choi, Heehyeon Kim, and Joyce Jiyoung Whang. 2025. Unveiling the threat of fraud gangs to graph neural networks: Multi-target graph injection attacks against gnn-based fraud detectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16028–16036.

Mingjiang Duan, Tongya Zheng, Yang Gao, Gang Wang, Zunlei Feng, and Xinyu Wang. 2024. Dga-gnn: Dynamic grouping aggregation gnn for fraud detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11820–11828.

Candida Maria Greco, Lucio La Cava, Lorenzo Zangari, and Andrea Tagarelli. 2025. Exploring llms' ability to spontaneously and conditionally modify moral expressions through text manipulation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18047–18070.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.

Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. Synergizing llms with global label propagation for multimodal fake news detection. *arXiv preprint arXiv:2506.00488*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

TN Kipf. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Mingxin Li, Yuchen Zhang, Haowei Xu, Xianghua Li, Chao Gao, and Zhen Wang. 2025. Learning complex heterogeneous multimodal fake news via social latent network inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 433–441.

Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. 2024. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2025. Help me write a story: Evaluating llms' ability to generate writing feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25827–25847.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2020. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv preprint arXiv:2009.03509*.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

S Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, and 1 others. 2023. Findings of factify 2: multimodal fake news detection. *arXiv preprint arXiv:2307.10475*.

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018a. 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*, pages 849–857.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018b. 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*, pages 849–857.

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*, pages 540–547.

Siqi Wei and Bin Wu. 2024. Intra and inter-modality incongruity modeling and adversarial contrastive learning for multimodal fake news detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 666–674.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2025. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843.

Chaowei Zhang, Zongling Feng, Zewei Zhang, Jipeng Qiang, Guandong Xu, and Yun Li. 2025a. Is llms hallucination usable? llm-based negative reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1031–1039.

Tianlin Zhang, En Yu, Yi Shao, and Jiande Sun. 2025b. Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection. *arXiv preprint arXiv:2502.01699*.

Zhongjian Zhang, Huabin Yang, Chenghao Huang, and Yanru Zhang. 2023. Team noir at factify 2: Multimodal fake news detection with pre-trained clip and fusion network. In *DE-FACTIFY@ AAAI*.

Wanqing Zhao, Yuta Nakashima, Haiyuan Chen, and Noboru Babaguchi. 2023. Enhancing fake news detection in social media via label propagation on cross-modal tweet graph. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2400–2408.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CS-02-108, Carnegie Mellon University*.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International conference on social informatics*, pages 109–123. Springer.

A Data Preprocessing

The data preprocessing pipeline focuses on multimodal LLM input adaptation and data quality assurance, covering two types of data: images and text.

Image Processing:Images undergo three-level standardized processing (scaling, padding, and fault tolerance). After successful reading, they are converted to RGB format and scaled to a maximum side length of 256pt while preserving the original aspect ratio. Blank areas are filled with white and aligned centrally. If image reading fails, a white-filled image of the target size is returned.

Text Processing:Text data is first subjected to data cleaning, followed by length truncation with a 1024-character limit.

B Prompt Design

Different prompt design methods are shown in Tables 3, 4, and 5.

C Training Strategy

The training strategy aims for low GPU memory usage and high training stability, with core adoption of 4-bit quantization and LoRA fine-tuning, covering the entire process of model initialization, optimizer configuration, training control, and model saving.

The base model is Qwen2.5-VL-7B, with LoRA configured as follows: low-rank dimension of 16, scaling factor of 32, and dropout rate of 0.05. The AdamW optimizer is employed, only updating LoRA parameters, with a learning rate of 2e-4 and a batch size of 4.

All experiments were conducted in a Python 3.10 environment equipped with an NVIDIA RTX 4090 GPU and 50 GB of memory. Training consisted of 5 epochs, with an early stopping strategy enabled: a validation was triggered every 100 iterations to save the optimal parameters; training terminated if no parameter updates were performed for 5 consecutive iterations.

D Baseline Model Details

- EANN (Wang et al., 2018b): Proposes an event-adversarial neural network framework for fake news detection, integrating textual and visual features via attention mechanisms. Its core lies in learning event-invariant features through an event discriminator and gradient reversal layer, with a focus on attention-driven multimodal fusion to enhance prediction accuracy and generalization across different events.

- SpotFake (Singhal et al., 2019): Targets fake news identification by leveraging comprehensive multimodal information, involving in-depth analysis of both textual content and accompanying images. It achieves effective detection performance by refining the feature alignment between text and images, ensuring coherent integration of cross-modal semantic information.

- MVAE (Khattar et al., 2019): A multimodal variational autoencoder designed for end-to-end fake news classification. It models the joint probability distribution of textual and image data, learning shared cross-modal representations through collaborative training of encoders, decoders, and a fake news detector. This integration of text and visual information optimizes the model's ability to distinguish between real and fake content.

- MCAN (Wu et al., 2021): A multimodal contextual attention network that enhances fake news detection by capturing both inter-modal and intra-modal relationships. It employs multi-layer attention mechanisms for iterative contextual modeling, effectively modeling dependencies between textual and image modalities to refine feature interactions.

- SAFE (Zhou et al., 2020): A similarity-aware multimodal model that combines feature extraction with cross-modal similarity measurement. It first extracts latent representations of text and images, then quantifies the similarity between modalities (e.g., via cosine similarity) to identify mismatches. This joint learn-

ing of tweet representations and inter-modal similarity enables accurate fake news detection.

- HMCAN (Qian et al., 2021): Constructs a hierarchical multimodal contextual attention network to capture rich hierarchical semantics for fake news detection. It uses BERT and ResNet for text and image encoding, respectively, and incorporates a hierarchical coding module to model high-level and fine-grained relationships, strengthening the model's understanding of complex multimodal content.

- FCN (Zhao et al., 2023): Relies on CLIP for unified multimodal feature extraction, constructing a cross-modal tweet graph to align textual and image features. It leverages graph convolutional networks (GCNs) to model structural correlations within the graph, thereby completing the fake news classification task.

- FCN-LP (Zhao et al., 2023): Builds on FCN's architecture with CLIP-based feature extraction and cross-modal tweet graph construction. It introduces a fixed iterative label propagation mechanism, which optimizes prediction results by propagating label information across graph nodes, further enhancing the model's discriminative performance.

- GLPN-LLM (Hu et al., 2025): An LLM-enhanced label propagation model for multimodal fake news detection. By integrating LLM-generated pseudo-labels with graph-based label propagation, it fuses strong semantic understanding from LLMs and structural information from graphs to optimize detection performance.

E Efficiency

The time complexity of the framework is analyzed as follows, focusing on conciseness and core focus:

- **LLM Pseudo-label Generation (based on Qwen2.5-VL-7B)**: Serves as the dominant complexity source of the framework. All samples undergo pseudo-label generation via the model. Let C_{LLM} denote the time complexity of a single multimodal inference of Qwen2.5-VL-7B. With N representing the total number of samples, the total complexity of this module is $O(N \cdot C_{\text{LLM}})$.

- **Cross-modal Label Propagation**: Consists of two steps: multimodal feature transformation and alignment, and K_{iter}-iterative label propagation. The total complexity is $O(N \cdot d^2 + K_{\text{iter}} \cdot M_{\text{edge}})$, where d is the multimodal feature dimension and M_{edge} is the number of edges in the cross-modal graph. Since C_{LLM} is significantly larger than the complexity of this module (large-model inference constitutes the main computational overhead), this part is negligible.

Overall, the time complexity of the framework is ultimately simplified to $O(N \cdot C_{\text{LLM}})$. By optimizing the prompt length and adjusting image resolution to reduce the actual computational cost of C_{LLM}, the framework can efficiently adapt to large-scale sample scenarios.

Table 3: Design of Text-unimodal Prompt in the Training Phase

Role	Content
system	You are a misinformation evaluator. Assess the news authenticity by the provided text: Label 1 if clearly true; 0 if ambiguous/unverifiable/suspicious. Output format: Strictly output only one line in the format "Result: R" (R=1/0). No additional text is allowed. Do not output any content beyond this line.
user	{text}
assistant	Result: {label}

Table 4: Design of Text-Image Multimodal Prompt in the Training Phase

Role	Content
system	You are a misinformation evaluator. Assess the news authenticity by text and image: Label 1 if clearly true; 0 if ambiguous/unverifiable/suspicious. Output format: Strictly output only one line in the format "Result: R" (R=1/0). No additional text is allowed. Do not output any content beyond this line.
user	{text} and {image}
assistant	Result: {label}

Table 5: Design of Text-image Multimodal Prompt in the Testing Phase

Role	Content
system	You are a misinformation evaluator. Assess the news authenticity by text and image: Label 1 if clearly true; 0 if ambiguous/unverifiable/suspicious. Provide a confidence score (0.1-1.0): lower (0.1-0.5) for unclear signals, higher (0.6-1.0) for certainty. Output in strict format (2 lines only, no extra text): - Line 1: "Result: R, Confidence: C" (R = 1/0, C=score). - Line 2: "Reason: E" (E = specific basis for judgment, which must simultaneously refer to text and image information).
user	{text} and {image}
assistant (inference)	Result: {label}, Confidence: {confidence} Reason: {explanation}