CIL: Causal-Intervention and LLM-Synergized Multimodal Fake News Detection

Hongjie Wang[1], Hongyu Chen[1], Peilin Jiang[1]*
[1] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
pljiang@xjtu.edu.cn

Abstract

Multimodal fake news detection requires models to reason over both textual and visual evidence while remaining robust to distribution shifts in social media graphs. Existing approaches often concatenate modality-specific representations or rely on post-hoc fusion, which can leave image and text features in misaligned semantic spaces. Graph-based methods further introduce structural bias: node popularity, degree, or connectivity may become shortcut signals that correlate with labels in the training distribution but do not causally determine veracity. We present CIL, a causal-intervention and LLM-synergized framework for multimodal fake news detection. CIL first uses CLIP to obtain aligned image and text representations and constructs a multi-view semantic graph from cross-modal, intra-modal, and fused similarities. It then uses a large language model to produce high-confidence pseudo labels, which are injected as weak supervisory anchors. Finally, CIL mitigates degree-induced structural confounding with inverse propensity weighting and enforces counterfactual consistency under edge-drop interventions. Experiments on Twitter, PHEME, and Weibo suggest that each component contributes to performance, with the largest gains coming from aligned multimodal representations and causal debiasing.

1 Introduction

The proliferation of multimodal fake news—false stories accompanied by misleading images—has severely threatened information credibility and public trust. Compared with traditional textual misinformation, multimodal fake news is significantly more deceptive because visual content can substantially enhance the credibility and emotional influence of misleading claims. In many cases, the attached images are partially authentic but semantically mismatched with the textual narrative, making unimodal verification insufficient. Detecting such disinformation is therefore highly challenging, as models must jointly reason over textual semantics, visual evidence, and social propagation structures while remaining robust to distribution shifts and noisy environments.

Another major challenge arises from the dynamic nature of social media platforms. News events evolve rapidly across domains and platforms, causing substantial discrepancies between training and testing distributions. Models trained on historical events often overfit event-specific semantics or platform-dependent propagation patterns, resulting in degraded generalization performance on unseen events. Moreover, fake news propagation is strongly influenced by popularity-related factors such as user engagement and network connectivity, which may induce spurious correlations unrelated to the actual authenticity of the content.

Existing multimodal fake news detection methods mainly follow two research directions: multimodal fusion models and graph-based reasoning models. Early multimodal approaches typically extract visual and textual features independently using encoders such as ResNet and BERT, followed by feature concatenation or attention-based fusion (Wang et al., 2018; Zhou et al., 2020; Singhal et al., 2019). Although these methods achieve promising performance, the independently pretrained encoders often produce representations residing in inconsistent semantic spaces, weakening cross-modal interaction and reducing semantic alignment quality.

Graph-based methods further improve fake news detection by modeling propagation structures or semantic relations among posts. A recent survey (Lakzaei et al., 2024), summarizes how GNN-based detectors exploit propagation graphs, similarity graphs, and heterogeneous social structures for disinformation detection. However, existing graph neural network methods implicitly assume

*Corresponding author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 405–414
November 6-10, 2026 ©2026 Association for Computational Linguistics

that graph topology faithfully reflects semantic relevance. In practice, graph structures in social media frequently contain severe structural bias. Node degree, popularity, and connectivity may become shortcut signals that correlate with labels in the training distribution but do not causally determine authenticity. Consequently, models may incorrectly rely on propagation statistics instead of intrinsic semantic evidence, leading to poor robustness under distribution shifts.

Recent advances in large language models (LLMs) provide new opportunities for weak supervision and knowledge-enhanced reasoning (Hu et al., 2024). Through prompt engineering and in-context learning, LLMs can generate informative pseudo-labels even in low-resource scenarios. Nevertheless, directly injecting LLM outputs into graph learning frameworks may introduce noisy supervision due to hallucinated or uncertain predictions, which can further propagate through graph structures and amplify classification errors.

To address these challenges, we propose CIL, a Causal-Intervention and LLM-Synergized framework for multimodal fake news detection. The proposed framework jointly integrates semantic alignment, graph reasoning, LLM-based weak supervision, and causal debiasing into a unified architecture.

Specifically, we first employ CLIP to project textual and visual information into a shared embedding space, thereby improving semantic consistency across modalities. Based on these aligned representations, we construct a multi-view semantic interaction graph that captures both intra-modal and cross-modal relationships among news samples. Second, we introduce LLM-generated pseudo-labels as weak supervisory signals. Instead of directly trusting all generated labels, we adopt a confidence-aware filtering strategy that retains only high-confidence predictions, reducing the influence of hallucinated outputs. More importantly, we explicitly model graph structural bias from a causal perspective. We identify node degree and popularity as potential confounding variables that create spurious correlations between graph topology and authenticity labels. To mitigate this issue, we incorporate inverse propensity weighting (IPW) and counterfactual consistency regularization, enabling the model to learn structure-invariant semantic representations and improving robustness under graph perturbations.

Extensive experiments on Twitter, PHEME, and Weibo datasets demonstrate that CIL consistently outperforms state-of-the-art methods under both multimodal and graph-based settings. Ablation studies further verify the effectiveness of semantic alignment, pseudo-label injection, and causal regularization components.

Our contributions are summarized as follows:

- We propose a unified multimodal fake news detection framework that combines semantic alignment, LLM-supervised weak learning, and causal regularization, effectively mitigating both semantic mismatch and structural bias.
- We design a multi-view semantic graph construction module using five complementary similarity measures, enabling more reliable relation modeling than conventional single-view graph construction strategies.
- We introduce causal intervention and counterfactual consistency regularization to reduce structural confounding and improve model robustness under graph perturbation and distribution shift.
- Extensive experiments on three real-world multimodal fake news datasets demonstrate the superiority of the proposed framework over competitive baselines and validate the effectiveness of each component through comprehensive ablation studies.

2 Related Work

2.1 Multimodal Fake News Detection

Multimodal fake news detection aims to identify misinformation by jointly modeling textual and visual content on social media platforms (Singhal et al., 2019; Khattar et al., 2019; Ying et al., 2023). Early methods mainly perform feature-level fusion by combining text encoders with visual backbones. For example, SpotFake (Singhal et al., 2019) and SAFE (Zhou et al., 2020) integrate BERT and ResNet features for multimodal classification and semantic consistency modeling. Later studies further explore fine-grained cross-modal interactions. MCAN (Wu et al., 2021) and HMCAN (Qian et al., 2021) use co-attention or hierarchical attention to capture correlations between textual tokens and image regions, while EANN (Wang et al., 2018) introduces adversarial learning for event-invariant representation learning.

Although effective, most existing methods rely on separately pretrained unimodal encoders and mainly focus on instance-level fusion, which may limit cross-modal semantic alignment and overlook latent relationships among news samples. In contrast, our framework adopts CLIP (Radford et al., 2021) to map images and texts into a shared semantic space, and further incorporates graph-based relational reasoning, following recent efforts such as MMDFND (Tong et al., 2024).

2.2 Graph-Based Fake News Detection

Graph neural networks (GNNs) have been widely used for fake news detection by modeling propagation structures, user interactions, and semantic relations among posts. FCN (Liu and Wu, 2018) applies graph convolution to rumor propagation modeling, while BiGCN (Bian et al., 2020) captures bidirectional propagation patterns. Later studies further construct heterogeneous graphs over users, posts, and events (Zhang et al., 2024). Recently, GLPN (Hu et al., 2025) combines graph propagation with LLM-generated pseudo-labels for low-resource fake news detection.

Despite their effectiveness, existing graph-based methods often assume that graph topology reliably reflects semantic relevance. In practice, social graphs can be affected by popularity bias, degree imbalance, and event-specific propagation patterns, leading models to exploit structural shortcuts rather than semantic evidence. Our framework addresses this issue through structural debiasing and counterfactual regularization.

2.3 Causal Debiasing and Counterfactual Learning

Causal inference has recently attracted increasing attention in machine learning due to its capability to distinguish genuine causal relationships from spurious correlations. Existing studies apply causal intervention and counterfactual reasoning to recommendation systems, computer vision, and natural language processing tasks.

Inverse propensity weighting (IPW) (Rosenbaum and Rubin, 1983) is a widely adopted strategy for mitigating confounding bias by constructing reweighted training distributions. Counterfactual learning further improves robustness by enforcing prediction consistency under perturbed environments or interventions. Several recent works apply causal learning to graph neural networks by learning causally invariant graph representations (Chen et al., 2022) or disentangling causal substructures from spurious bias factors (Fan et al., 2022).

Nevertheless, causal debiasing remains largely underexplored in multimodal fake news detection. Existing fake news detection methods rarely consider the causal effects of graph structure and propagation bias, especially in combination with multimodal alignment and LLM-based supervision. To the best of our knowledge, our work is the first to jointly integrate causal intervention, counterfactual reasoning, multimodal semantic alignment, and LLM-driven weak supervision into a unified fake news detection framework.

2.4 LLM in Weak Supervision

Large language models (LLMs) have demonstrated remarkable reasoning and knowledge acquisition capabilities across a wide range of NLP tasks (Touvron et al., 2023; OpenAI et al., 2024). Recent studies increasingly utilize LLMs for weak supervision, pseudo-label generation, and low-resource learning. Through prompt engineering and in-context learning, LLMs can provide informative annotations even without task-specific fine-tuning (Dong et al., 2024; Brown et al., 2020).

Several recent fake news detection methods introduce LLMs as external knowledge providers to improve generalization and reduce annotation costs (Hu et al., 2024). However, LLM-generated labels are not always reliable and may contain hallucinated or inconsistent predictions. Existing approaches typically adopt simple confidence filtering or directly utilize raw outputs, which may propagate noisy supervision signals into downstream graph learning models.

Different from previous methods, our framework combines confidence-aware pseudo-label selection with causal regularization. By retaining only high-confidence pseudo-labels and introducing counterfactual consistency constraints, the proposed framework effectively reduces the negative influence of noisy supervision while improving robustness and generalization.

3 Methodology

3.1 Overview

We propose CIL, as illustrated in Fig. 1, a unified framework for multimodal fake news detection that consists of three modules: semantic-aligned graph construction, LLM-driven pseudo-label injection, and causal debiasing. Given a news instance with

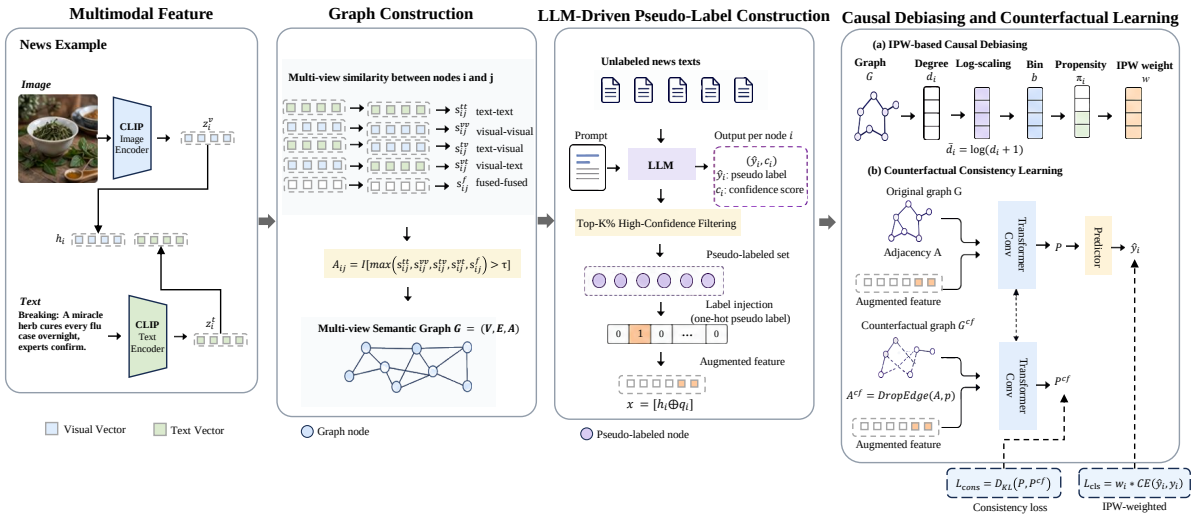


Figure 1: Overview of the proposed CIL framework.

text and image, CIL first uses CLIP to encode both modalities into a shared semantic space, where multi-view similarities are computed to construct a semantic graph among news samples. Then, an LLM is used to generate confidence-filtered pseudo-labels for unlabeled training samples, providing additional weak supervision under limited annotation settings. Finally, the constructed graph and enhanced node representations are fed into a graph neural network, where inverse propensity weighting and counterfactual consistency regularization are applied to reduce structural bias caused by degree imbalance and spurious graph patterns. The model outputs the predicted fake-news label for each news instance.

3.2 Semantic-Aligned Multimodal Graph Construction

Textual and visual contents in social media posts usually lie in heterogeneous feature spaces, making direct feature concatenation insufficient for semantic alignment. To obtain modality-consistent representations, we adopt CLIP (Radford et al., 2021) as the backbone instead of separately using conventional text and image encoders. Given a dataset $D = \{P_i\}_{i=1}^{N}$, each news sample $P_i = (T_i, V_i)$ contains textual content T_i and visual content V_i. We encode them with the CLIP text encoder E_{text} and image encoder E_{img}, and concatenate the resulting features as the initial node representation: $\mathbf{t}_i = E_{\text{text}}(T_i)$, $\mathbf{v}_i = E_{\text{img}}(V_i)$, and $\mathbf{x}_i = \mathbf{t}_i \oplus \mathbf{v}_i$, where $\mathbf{t}_i, \mathbf{v}_i \in \mathbb{R}^d$ and $\mathbf{x}_i \in \mathbb{R}^{2d}$.

Based on these aligned representations, we construct a multi-view semantic graph to capture latent relations among news samples. For each node pair (i, j), we compute cosine similarities from five views, including text-text, image-image, image-text, text-image, and concatenated feature similarity, where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}/(\|\mathbf{u}\|\|\mathbf{v}\|)$. An edge is established when the maximum similarity among these views exceeds a threshold θ: $A_{ij} = \mathbb{I}(\max(S_{\text{cat}}, S_{i2t}, S_{t2i}, S_{i2i}, S_{t2t}) > \theta)$. This construction captures both intra-modal homophily and cross-modal semantic complementarity while keeping the graph sparse for subsequent graph learning.

3.3 LLM-Driven Pseudo-Label Injection

Since labeled fake news data are often limited, we use LLMs as external weak annotators to provide additional supervision. Specifically, we design a Chain-of-Thought prompting template that asks the LLM to act as a misinformation assessment expert and evaluate each news sample according to predefined criteria. The detailed prompt is shown in Table 1. For each unlabeled training sample u, we parse the LLM output to obtain a predicted label $\hat{y}_u$ and a confidence score c_u. To reduce noisy supervision caused by hallucinated or unreliable outputs, only samples with confidence scores in the top $K\%$ are retained as high-confidence pseudo-labeled samples:

$$D_{\text{pseudo}} = \left\{ (x_u, \hat{y}_u) \mid c_u \geq \text{Percentile}_{(1-K)}(c) \right\}. \quad (1)$$

During training, these pseudo-labels are used as soft anchors and injected into the node feature space. The original node feature is expanded from $2d$ to $2d + 2$, where the additional two dimensions correspond to the binary label space. For

System Prompt: You are a professional misinformation assessment expert with extensive experience in fake news detection and evaluation. Your primary task is to assess the authenticity of the provided news content. You must strictly follow these evaluation criteria:
(1) If the news is clearly true, mark it as **1 (True)**.
(2) If the news contains ambiguous, unverifiable information or questionable claims, you must classify it as **0 (False)**.
(3) Along with classification, provide a confidence score (0% - 100%) reflecting your certainty in this decision.
(3.1) When the news contains unclear or mixed signals, the confidence should be low (e.g., 50%-70%);
(3.2) Only give high confidence (e.g., 80%-100%) when you are very certain.
Output Format Requirement: Result: R, Confidence: C%, where R is 1 or 0, and C is the confidence value. Please be precise, concise, and provide a brief justification.
User Input: <CLS><Prompt><SEP><Cleaned News Text>

Table 1: Prompt design for fake news detection task

labeled training samples and samples in D_{pseudo}, these dimensions are filled with the corresponding one-hot label vectors; for other unlabeled samples, they are set to zero. In this way, high-confidence LLM predictions provide auxiliary supervision for graph learning while reducing the influence of noisy pseudo-labels.

3.4 Causal Debiasing and Counterfactual Reasoning

Although graph structures provide useful relational information, they may also introduce structural bias into fake news detection. We formulate this problem from a causal perspective with three variables: multimodal content X, graph structure A, and authenticity label Y. Ideally, the model should predict Y mainly from content-related evidence. However, graph learning may be affected by a confounder C, which is mainly reflected by node degree centrality. Since highly connected nodes are more likely to form dense neighborhoods and may also correlate with specific labels, a backdoor path $A \leftarrow C \rightarrow Y$ can arise. As a result, conventional graph neural networks may learn $P(Y|A, X)$ by exploiting degree-related shortcuts rather than reliable semantic evidence.

To reduce this effect, we adopt inverse propensity weighting (IPW) to reweight training samples according to their degree-related structural bias. Specifically, we compute the degree d_i of each node and apply logarithmic smoothing:

$$c_i = \log(d_i + 1). \tag{2}$$

The smoothed degree values are discretized into M bins, and $B(i)$ denotes the bin index of node i. The propensity score is estimated by the proportion of training samples in the corresponding bin:

$$e_i \approx \frac{N_{\text{train}}^{(B(i))}}{N_{\text{all}}^{(B(i))}}. \tag{3}$$

The training weight is then defined as $w_i = 1/e_i$, giving larger weights to underrepresented structural patterns and reducing the model's reliance on degree-related shortcuts.

Beyond sample-level reweighting, we further introduce counterfactual graph perturbation to improve robustness against structural changes. Given the original adjacency matrix A, we construct a perturbed graph by random edge dropout, i.e., $\tilde{A} = \text{Dropout}(A, p)$. The original graph (X, A) and the perturbed graph $(X, \tilde{A})$ are fed into the model to obtain prediction distributions Q and $\tilde{Q}$, respectively. We then enforce prediction consistency through a KL-divergence regularization term:

$$L_{\text{consist}} = \frac{1}{|V|} \sum_{i \in V} D_{\text{KL}}(Q_i \parallel \tilde{Q}_i) \tag{4}$$

This objective is applied to all nodes, including unlabeled ones, encouraging the model to learn representations that remain stable under graph perturbations.

As the graph aggregation backbone, we use a Transformer-based multi-head attention layer instead of a conventional GCN. At the l-th layer, the node representation is updated as:

$$\mathbf{h}_i^{(l)} = \text{Concat}_{k=1}^{H} \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k W_V^k \mathbf{h}_j^{(l-1)} \right), \tag{5}$$

where the attention coefficient is computed by Query-Key dot-product attention:

$$\alpha_{ij}^k = \text{softmax}_j \left(\frac{\left(W_Q^k \mathbf{h}_i^{(l-1)}\right)^\top \left(W_K^k \mathbf{h}_j^{(l-1)}\right)}{\sqrt{d_{\text{head}}}} \right). \tag{6}$$

This attention mechanism allows the model to adaptively aggregate neighborhood information according to semantic relevance.

3.4.1 Optimization Objective

The final loss function consists of two components, namely the causally weighted classification loss L_{class} and the counterfactual consistency loss L_{consist}:

$$L_{\text{class}} = -\frac{1}{|D_{\text{train}}|} \sum_{i \in D_{\text{train}}} w_i \times \sum_{c=1}^{C} y_{ic} \log P(y_{ic} \mid x_i, A). \tag{7}$$

where w_i denotes the IPW weight computed previously. The overall optimization objective is formulated as:

$$\min_{\Theta} L_{\text{total}} = L_{\text{class}} + \lambda \cdot L_{\text{consist}} \tag{8}$$

where λ is a balancing coefficient that controls the strength of the consistency regularization term.

3.4.2 Complexity Analysis

We further analyze the computational complexity of the proposed CIL model. Assume that the graph contains N nodes and E edges, with feature dimension D. During the graph construction stage, the similarity computation requires a complexity of $O(N^2D)$. However, this operation belongs to the offline preprocessing stage and therefore does not contribute to the training overhead.

During the IPW computation stage, only degree statistics and binning operations are involved, resulting in a complexity of $O(N)$. During model training, the single-layer complexity of the feature aggregation network is $O(ED + ND^2)$. Since counterfactual intervention is introduced, each iteration requires two forward propagations, corresponding to the original graph and the intervened graph, respectively. Therefore, the total training complexity can be approximated as:

$$2 \times L \times O(ED + ND^2) \tag{9}$$

where L denotes the number of graph neural network layers.

The analysis demonstrates that although the proposed framework introduces sophisticated causal debiasing mechanisms, the training complexity remains linearly scalable with respect to the graph structure, maintaining computational efficiency comparable to standard graph neural networks and exhibiting strong scalability for large-scale social network data.

4 Experiments

4.1 Datasets and Metrics

We evaluate CIL on three widely used multimodal fake news detection datasets: Twitter (Boididou et al., 2015), PHEME (Zubiaga et al., 2016), and Weibo (Jin et al., 2017), which cover different platforms, languages, and event types. Twitter contains 17,441 English tweets with images from major social events, PHEME includes about 2,400 event-related tweets with propagation structures, and Weibo contains 5,214 Chinese image-text posts from domains such as disasters, healthcare, economics, and military affairs.

Following previous setting (Hu et al., 2025), each dataset is randomly split into training, validation, and test sets with a ratio of 70%/10%/20%. To simulate low-resource settings, only 5% of the training samples are labeled, while the remaining training samples are used as unlabeled data for graph propagation and pseudo-label learning. All experiments are repeated five times with different random splits, and average results are reported. We use Accuracy, Precision, Recall, and F1-score as evaluation metrics, with F1-score as the primary metric.

4.2 Experimental Setup

CIL is implemented with PyTorch 2.0. We use CLIP ViT-B/32 as the multimodal encoder, producing 1024-dimensional image-text features, which are expanded to 1026 dimensions after pseudo-label injection. For graph construction, five cosine similarities are used with threshold $\theta = 0.70$, and each node keeps at least three nearest neighbors to avoid isolation.

GPT-4o generates pseudo-labels, and only the top 5% high-confidence predictions are retained. The graph encoder is a two-layer TransformerConv network with eight attention heads, followed by LayerNorm and Dropout ($p = 0.3$) after the first layer. For causal debiasing, node degrees are log-transformed and discretized into 10 bins for propensity estimation, with IPW weights clipped to $[0.1, 10]$. Counterfactual learning uses edge dropout with probability $p = 0.5$ and KL-divergence consistency regularization.

Method	Twitter				PHEME				Weibo			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
LLM (2024)	75.39	75.66	80.92	78.20	74.38	78.66	75.16	76.87	80.86	82.16	81.33	81.75
EANN (2018)	71.53	71.38	63.82	68.91	70.17	71.28	67.36	69.10	79.18	80.31	78.52	79.44
MVAE (2019)	74.56	80.15	76.34	81.57	77.83	73.82	73.45	72.21	71.86	70.32	70.32	70.53
SAFE (2020)	76.66	76.32	75.41	76.37	81.25	79.22	79.11	79.69	84.91	83.81	82.19	83.01
SpotFake (2019)	77.16	75.32	87.83	85.14	81.37	79.53	81.22	79.43	86.39	86.12	87.17	83.22
MCAN (2021)	80.91	82.68	76.67	82.26	80.74	79.21	79.64	80.15	86.50	88.10	84.60	86.15
HMCAN (2021)	83.91	81.68	84.67	82.57	86.36	83.18	83.81	83.49	86.75	88.40	84.65	87.20
FCN (2018)	82.86	78.64	87.39	82.78	80.36	84.43	89.12	86.71	82.92	83.17	88.45	86.74
GLPN (2025)	87.60	86.52	88.88	86.86	88.29	86.92	88.14	86.87	**90.66**	89.06	92.20	91.46
CIL (Ours)	**90.67**	**87.12**	**90.23**	**88.65**	**88.34**	**89.67**	**92.01**	**90.82**	89.45	**90.78**	**94.12**	**92.42**

Table 2: Main experimental results across three multimodal fake news detection datasets

The model is trained with Adam using a learning rate of 1×10^{-4} and weight decay of 5×10^{-4}. Early stopping is applied if validation F1 does not improve for 20 epochs. All experiments are conducted on an NVIDIA RTX 4090 GPU. Baselines follow the same splits and evaluation protocols whenever possible, with hyperparameters selected from their original papers. For LLM-based baselines, the same GPT-4o model and prompting strategy are used.

4.3 Baselines

We compare CIL with representative multimodal fake news detection methods. The baselines include an LLM-only classifier (Hu et al., 2024), EANN (Wang et al., 2018), MVAE (Khattar et al., 2019), SAFE (Zhou et al., 2020), SpotFake (Singhal et al., 2019), MCAN (Wu et al., 2021), HMCAN (Qian et al., 2021), FCN (Liu and Wu, 2018), and GLPN (Hu et al., 2025). These methods cover adversarial event-invariant learning, multimodal variational modeling, similarity-aware fusion, dual-stream multimodal modeling, co-attention, hierarchical attention, graph-based label propagation, and LLM-enhanced pseudo-labeling. GLPN is the most direct baseline, as it also uses LLM-derived pseudo labels; CIL differs by adding CLIP-aligned multi-view graph construction and causal structural debiasing.

4.3.1 Main Experimental Results

Table 2 presents the performance comparison between the proposed CIL model and several baseline methods on the three datasets. Overall, CIL achieves the best or near-best performance across most evaluation metrics, demonstrating the effectiveness of the proposed framework.

On the Twitter dataset, CIL achieves an accuracy of 90.67% and an F1-score of 88.65%, outperforming GLPN by 3.07% and 1.79%, respectively. This improvement demonstrates the effectiveness of the proposed causal intervention and multi-view graph construction strategies. By removing structural bias through IPW and enhancing robustness via counterfactual learning, the model can focus more on semantic content rather than popularity-related features.

On the PHEME dataset, CIL achieves the best F1-score of 90.82%, outperforming GLPN by 3.95%. Since PHEME is relatively small and contains limited event categories, the results indicate that CIL maintains strong generalization ability under low-resource settings. The improvement mainly benefits from high-confidence pseudo-labels generated by the LLM and the semi-supervised graph propagation mechanism.

On the Weibo dataset, CIL achieves an F1-score of 92.42%, which is higher than GLPN by 0.96%. This result demonstrates that the proposed framework also exhibits strong cross-language and cross-domain transferability. The semantic alignment capability of CLIP and the robustness introduced by causal debiasing both contribute to this improvement.

Further analysis shows that CIL maintains a better balance between Precision and Recall compared with baseline methods. For example, on the Twitter dataset, CIL achieves 87.12% Precision and 90.23% Recall, indicating that the model can accurately identify fake news while maintaining high coverage. In contrast, several baseline methods exhibit more obvious imbalance between the two metrics.

4.4 Ablation Study

To evaluate the contribution of each component in CIL, we conduct ablation experiments on the Twitter dataset. The results are shown in Table 3.

Variant	Acc	Pre	Rec	F1
CIL	**90.67**	**87.12**	**90.23**	**88.65**
w/o CLIP	88.45	84.89	88.01	86.42
w/o MV Graph	89.23	86.67	88.78	87.71
w/o LLM PL	89.56	87.01	89.12	88.05
w/o IPW	89.34	85.78	88.89	87.31
w/o CF Reg.	89.78	86.23	89.34	87.76
w/o IPW + CF	88.89	85.34	88.45	86.87

Table 3: Ablation study on the Twitter dataset (%).

Removing any module leads to performance degradation, demonstrating the effectiveness of each component. Replacing CLIP with ResNet+BERT causes the largest drop, showing the importance of cross-modal feature alignment. Removing the multi-view graph construction also reduces performance, indicating that multiple similarity views provide richer semantic relations.

Without LLM-generated pseudo-labels, the model loses useful external supervision under the semi-supervised setting. Removing IPW debiasing significantly decreases performance, verifying the importance of causal intervention in alleviating structural bias. In addition, counterfactual regularization improves robustness by enforcing prediction consistency under graph perturbations. When both IPW and counterfactual regularization are removed, the performance drops further, confirming that the two modules complement each other.

4.5 Hyperparameter Analysis

To further investigate the robustness of the proposed framework, we analyze the influence of several important hyperparameters on the Twitter dataset, including the consistency weight λ, edge dropout probability p, pseudo-label retention ratio K, and IPW bin number M. The results are summarized in Table 4.

For the consistency weight λ, the performance first improves and then declines as λ increases. The best performance is achieved at $\lambda = 0.3$, indicating that moderate consistency regularization improves robustness and generalization, while excessively large values overemphasize consistency constraints and weaken the classification objective.

For the edge dropout probability p, the best performance is obtained at $p = 0.5$. Moderate graph perturbation improves robustness against structural bias, whereas overly large dropout ratios damage graph connectivity and reduce propagation effectiveness.

Parameter	Setting	Acc	Pre	Rec	F1
λ	0.0	89.78	86.23	89.34	87.76
	0.1	90.12	86.56	89.67	88.09
	0.2	90.45	84.89	90.01	87.38
	0.3	**90.67**	**87.12**	**90.23**	**88.65**
	0.4	90.56	87.01	90.12	88.54
	0.5	90.34	86.78	89.89	88.31
p	0.0	89.78	86.23	89.34	87.76
	0.2	90.23	86.67	89.78	88.20
	0.3	90.45	86.89	90.01	88.42
	0.4	90.56	87.01	90.12	88.54
	0.5	**90.67**	**87.12**	**90.23**	**88.65**
	0.7	90.12	86.56	89.67	88.09
$K(\%)$	0	89.56	86.01	89.12	87.54
	1	89.89	86.34	89.45	87.87
	3	90.34	86.78	89.89	88.31
	5	**90.67**	**87.12**	**90.23**	**88.65**
	10	90.23	86.67	89.78	88.20
M	5	89.89	86.34	89.45	87.87
	8	90.34	86.78	89.89	88.31
	10	**90.67**	**87.12**	**90.23**	**88.65**
	15	90.45	86.89	90.01	88.42
	20	90.12	86.56	89.67	88.09

Table 4: Hyperparameter analysis on the Twitter dataset highlights key parameter effects. (%).

For the pseudo-label retention ratio K, the model achieves optimal performance at $K = 5\%$. A small number of high-confidence pseudo-labels provides effective external supervision, while larger ratios introduce noisy pseudo-labels and negatively affect performance.

Finally, for the IPW bin number M, the best result is obtained at $M = 10$. Small bin numbers lead to coarse propensity estimation, while excessively large values produce unstable estimation due to insufficient samples in each interval.

5 Conclusion

We propose CIL, a multimodal fake news detection framework that combines CLIP-based semantic alignment, LLM-assisted weak supervision, and structural debiasing. CIL constructs a multi-view semantic graph, injects confidence-filtered pseudo-labels, and applies degree-aware reweighting with counterfactual consistency regularization. Experiments on Twitter, PHEME, and Weibo demonstrate its effectiveness over competitive baselines, while ablation results confirm the contribution of key components.

Limitations

Although CIL shows strong empirical performance, several limitations remain. First, the quality of LLM-generated pseudo-labels depends on the capability and calibration of the external language model. Even with confidence filtering, incorrect high-confidence predictions may still be injected into the graph and affect downstream training. Second, the framework relies on CLIP to provide aligned multimodal representations. When images contain fine-grained visual evidence, manipulated regions, or domain-specific cues that are not well captured by the pretrained encoder, the constructed semantic graph may be less reliable. Third, the multi-view graph construction requires pairwise similarity computation among samples, which may become expensive for very large-scale social media streams despite being performed as an offline preprocessing step. Finally, the current evaluation focuses on three benchmark datasets. More experiments on real-time, cross-platform, and multilingual misinformation scenarios are needed to further validate the generalization ability of the proposed framework.

Ethics Statement

This work aims to support multimodal fake news detection and should be used as an assistive tool rather than a fully automated decision-making system. The datasets used in this study are existing research benchmarks, and we do not attempt to identify individual users or infer sensitive personal attributes. The framework uses LLM-generated pseudo-labels, which may contain errors or biases even after confidence filtering. Therefore, model predictions should be interpreted with caution, especially in high-stakes scenarios such as content moderation, journalism, public health, or political communication. Human review is necessary before taking real-world actions based on the model output. We also acknowledge that biased or incomplete training data may lead to unfair performance across topics, languages, or communities. Future deployment should include transparency, bias auditing, and regular cross-domain evaluation.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. 2022. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 22131–22148. Curran Associates, Inc.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing graph neural networks via learning disentangled causal substructure. In *Advances in Neural Information Processing Systems*, volume 35, pages 24934–24946. Curran Associates, Inc.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1440.

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 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Dhruv Khattar, Jaipal Singh, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational

autoencoder for fake news detection. pages 2915–2921.

Batool Lakzaei, Mostafa Haghir Chehreghani, and Alireza Bagheri. 2024. Disinformation detection using graph neural networks: a survey. *Artificial Intelligence Review*, 57(3):52.

Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

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. Association for Computational Linguistics.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 5384–5392.

Yuchen Zhang, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2024. Heterogeneous subgraph transformer for fake news detection. In *Proceedings of the ACM web conference 2024*, pages 1272–1282. Association for Computing Machinery.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: similarity-aware multi-modal fake news detection. *CoRR*, abs/2003.04981.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11.