CIAF: Cross-modal Inconsistency Aware Framework for Multimodal Fact-Checking

Zidong Yi[1]

[1] Collage of Software, Xi'an Jiaotong University, Xi'an, China
jfm200408@stu.xjtu.edu.cn

Abstract. In the multimodal information era, fact-checking faces growing challenges as misinformation becomes increasingly complex. Content that appears plausible in each individual modality may still contain subtle inconsistencies across modalities, which are often overlooked by traditional methods. Existing approaches mainly fuse multimodal features but rarely explicitly model fine-grained cross-modal inconsistencies, limiting both accuracy and interpretability. To address this issue, we propose a Cross-modal Inconsistency Aware Framework (CIAF) that explicitly identifies and localizes implicit discrepancies between textual claims and corresponding images. Specifically, CIAF first extracts fine-grained textual and visual representations and uses affective cues only as auxiliary evidence for discrepancy estimation. It then introduces a novel relational discrepancy attention (RDA) mechanism to model semantic and affective interactions between modalities, dynamically weighting potential inconsistency signals. Furthermore, we design an evidence-conditioned consistency reasoning module to aggregate localized discrepancy evidence for final fact-checking decisions, thereby enhancing both precision and robustness. Experiments on two challenging multimodal fact-checking benchmarks show that our approach delivers superior performance and interpretability, underscoring the importance of modeling fine-grained cross-modal inconsistencies for robust fact verification.

Keywords: Multimodal fact-checking, Cross-modal inconsistency modeling, Evidence consistency reasoning.

1 Introduction

Information dissemination on social media platforms such as Twitter, Facebook, and Weibo has grown explosively over the past decade. The massive influx of multimodal content—combining textual statements with visual media—has greatly increased communication efficiency and user engagement. However, this rapid proliferation of information has simultaneously amplified the potential harm caused by misleading claims and low-cost manipulated media, commonly referred to as *cheapfakes* [7]. In recent years, multimodal fact-checking approaches [11,24,3] have emerged to leverage the complementary information from both text and images. These approaches generally follow three major paradigms: (1) *cross-modal contrast*, which aligns representations

by contrasting matching and non-matching text-image pairs; (2) *cross-modal translation or alignment*, which translates or projects one modality into the space of the other to achieve semantic correspondence; and (3) *cross-modal fusion*, which integrates features from multiple modalities into a unified representation for downstream classification. While these paradigms have achieved notable performance improvements, they predominantly focus on global alignment or holistic fusion of representations. A particularly challenging aspect of multimodal misinformation lies in the strategic pairing of textual claims with images: while each modality alone may appear plausible, subtle inconsistencies across modalities often serve as critical cues for detecting falsity. For instance, semantic contradictions, entity-event mismatches, affective tone conflicts, or local visual anomalies can provide strong evidence that a multimodal claim is misleading or unsupported. Unfortunately, existing fact-checking approaches [17,18,25,20] frequently overlook these fine-grained cross-modal contradictions, which limits both the accuracy and interpretability of current models.

Recognizing these limitations, we propose the **Cross-modal Inconsistency Aware Framework (CIAF)**, which explicitly targets the detection and localization of fine-grained inconsistencies between textual claims and corresponding images. CIAF uses affective information only as auxiliary discrepancy evidence, while the core of the framework is explicit cross-modal inconsistency modeling and localization. The framework comprises three core components. First, it obtains fine-grained token- and patch-level representations for textual claims and visual evidence, and augments them with lightweight affective cues when such cues are available. Second, we introduce a novel **Relational Discrepancy Attention (RDA)** mechanism that captures semantic and emotional interactions at a fine-grained level across modalities. RDA dynamically assigns attention weights to regions and tokens that are likely to exhibit cross-modal contradictions, enabling precise localization of misleading content. Finally, CIAF incorporates an *evidence-conditioned consistency reasoning* module, which summarizes localized discrepancy evidence and guides the classifier toward targeted cross-modal consistency evaluation, facilitating improved generalization, particularly in low-resource scenarios.

We conduct extensive experiments on two challenging multimodal fact-checking benchmarks. The results show that CIAF consistently outperforms representative baseline models in both accuracy and F1 score. Further analyses reveal that explicitly modeling fine-grained cross-modal inconsistencies, together with auxiliary affective discrepancy cues, substantially improves the robustness and interpretability of multimodal fact verification systems.

2 Related Work

2.1 Multimodal Fact-Checking

Fact-checking seeks to assess the truthfulness of a given claim [7], playing a vital role in areas such as cheapfake detection, rumor verification, and claim verification. While early fact-checking approaches mainly focused on textual analysis [6,14,22], recent developments have increasingly integrated multimodal information, such as images

[11,26] and videos [3,13] to combine diverse sources of evidence for more reliable verification and justification.

Recent research on multimodal fact-checking (MFC) can be broadly categorized into three lines: cross-modal contrast methods [17,25,18,28], cross-modal translation methods [15,9], and cross-modal fusion methods [20,24,16].

Cross-modal contrast methods. These methods typically evaluate information veracity by measuring the similarity between different modalities, such as text and images, to detect inconsistencies. The underlying assumption is that multimodal misinformation often combines textual and visual elements that conflict to create a false sense of credibility. Building on this, some studies use cross-modal similarity as a weighting factor to balance unimodal and multimodal features, improving heterogeneous information fusion. For example, [17] proposed a multi-modal classifier ensemble that combines single- and multi-modal classifiers with center loss to enhance discriminative feature learning. [25] developed COOLANT, a cross-modal contrastive learning framework that aligns text and image features and uses attention to aggregate unimodal and cross-modal representations. [18] exploited inter-modality discordance with a modified contrastive loss to capture inconsistencies between text and visuals and [28] introduced Bootstrapping Multi-view Representations (BMR), which extracts separate representations from multiple views of text and image features, evaluates cross-modal consistency, and adaptively reweights each view to improve detection performance.

Cross-modal translation methods. The core idea of these methods is to bridge the semantic gap between modalities through transformation or alignment before classification. For example, image descriptions can be generated and aligned with textual content to form a unified semantic representation, which is then used by a classifier. [15] proposed DGExplain, an explainable multimodal framework for COVID-19 misinformation detection, explicitly modeling cross-modal associations and leveraging user comments for interpretability. Similarly, [9] enhanced image-text alignment by incorporating object features, textual features, and captions, applying boosting strategies to improve classification accuracy on social media news datasets.

Cross-modal fusion methods. These methods typically fuse multimodal features using architectures such as multilayer perceptrons, Transformers, or graph neural networks to obtain joint representations that capture both intra- and inter-modal relationships. Classification is then performed based on these fused features. For example, [20] fine-tuned state-of-the-art multimodal Transformers for claim verification and proposed data augmentation strategies to enhance generalization and robustness against manipulated social media inputs. [24] proposed Miner-uvs, which integrates unimodal veracity classifiers to weight features before fusion, mitigating conflicts across modalities and producing more discriminative representations. [16] introduced HAMMER and HAMMER++, using manipulation-aware contrastive learning and modality-aware cross-attention for fine-grained reasoning, enabling both detection of fake content and localization of manipulated regions in text and images.

Despite the progress of existing multimodal fact-checking methods, they generally share common limitations: most approaches either focus on global feature alignment or simple fusion, often overlooking fine-grained cross-modal inconsistencies and sentiment-related cues. This limits both verification accuracy and interpretability, motivating the need for models that can explicitly capture subtle discrepancies between textual claims and visual evidence.

2.2 Cross-Modal Inconsistency Modeling

Cross-modal consistency modeling captures semantic, emotional, and logical relationships across modalities (e.g., text–image, text–video) to support tasks such as multimodal retrieval, visual question answering, and fact-checking. Early approaches [12,29,4] focused on global semantic alignment by learning joint embedding spaces, such as CLIP, which assess correspondence between textual descriptions and visual content, but often miss fine-grained inconsistencies. Recent methods [10,21,31] address this by modeling region–word alignments and applying attention-based fusion to detect potentially conflicting fragments. Nonetheless, existing techniques remain limited in capturing fine-grained inconsistencies, integrating emotional cues with semantic information, and providing interpretable reasoning. To overcome these limitations, we propose the CIAF, which integrates emotional and semantic representations and leverages a Relational Discrepancy Attention mechanism to explicitly identify and localize inconsistencies between textual claims and visual evidence, improving both verification accuracy and explainability.

3 Proposed Method

In this section, we introduce the proposed **Cross-modal Inconsistency Aware Framework (CIAF)**. The core objective of CIAF is to enhance the accuracy and interpretability of multimodal fact-checking by moving beyond coarse-grained feature fusion to explicitly model and quantify fine-grained inconsistencies between text and image modalities. We posit that multimodal misinformation often manifests as subtle mismatches in semantic meaning, affective tone, entity-event correspondence, or logical coherence across modalities. By detecting these discrepancies, our framework not only improves classification performance but also provides interpretable cues for why a multimodal claim might be unsupported or misleading. The architecture, as illustrated in Fig. **1**, is composed of five synergistic modules: (1) multimodal feature representation, (2) auxiliary affective cue extraction, (3) cross-modal inconsistency modeling via Relational Discrepancy Attention (RDA), (4) evidence-conditioned consistency reasoning, and (5) task learning and decision making with a multi-task objective.

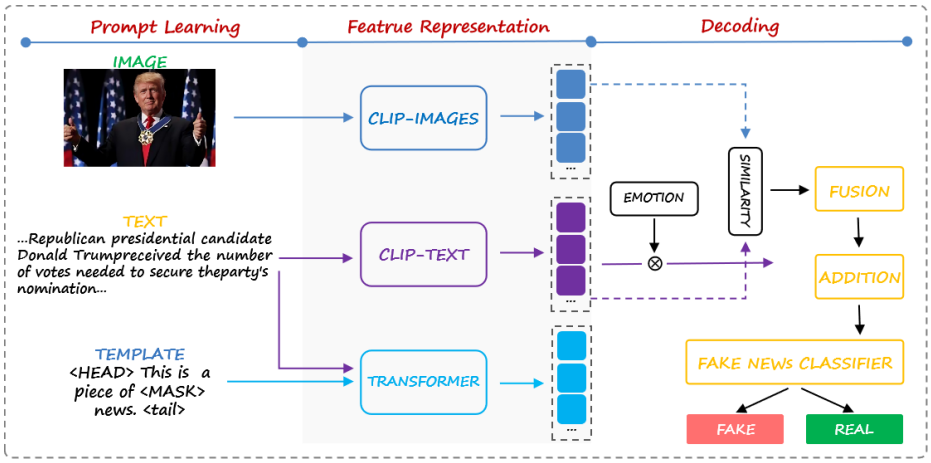

Fig. 1. The overall architecture of the proposed CIAF

3.1 Feature Representation

A robust joint representation space is foundational for cross-modal analysis. To this end, we leverage the **CLIP model** [12], renowned for its ability to align textual and visual concepts in a shared semantic embedding space through large-scale contrastive pre-training. For an input text claim $W = (w_1, w_2, \ldots, w_n)$ and its associated image I, CLIP processes them independently. The text encoder $f_T^{seq}(\cdot)$ and image encoder $f_V^{seq}(\cdot)$ produce sequences of token-level and patch-level embeddings, respectively:

$$T = f_T^{seq}(W) \in \mathcal{R}^{n \times d}, \quad V = f_V^{seq}(I) \in \mathcal{R}^{m \times d}, \tag{1}$$

where n and m denote the sequence lengths, and d is the shared embedding dimension. These token/patch-level features preserve fine-grained information crucial for later inconsistency detection. Additionally, we obtain global modality representations T_w and V_i via average pooling over the token and patch sequences, serving as holistic summaries for each modality.

3.2 Auxiliary Affective Cue Extraction

Affective information can provide useful auxiliary evidence for multimodal fact-checking. For example, a highly sensational claim paired with a visually neutral image may indicate a potential cross-modal mismatch. In CIAF, affective cues are not treated as the primary decision basis; instead, they are used as lightweight auxiliary signals to help estimate whether textual and visual evidence express inconsistent meanings or tones.

Textual Affective Cue Injection. For each text sample W, we employ a sentiment analysis tool (e.g., TextBlob) to extract two key dimensions: *polarity* $p \in [-1,1]$ (negative to positive) and *subjectivity* $s \in [0,1]$ (objective to subjective). These scores are first normalized to a consistent range:

$$\tilde{p} = \frac{p - p_{\min}}{p_{\max} - p_{\min}}, \quad \tilde{s} = s, \tag{2}$$

where $p_{\min}$ and $p_{\max}$ are dataset-specific minimum and maximum polarity values. A composite scalar emotional score e_T is computed as a weighted sum, $e_T = \lambda_1 \tilde{p} + \lambda_2 \tilde{s}$. This score is then projected to a d-dimensional vector $E_T = e_T \cdot \mathbf{1}_d$ and infused into the original token embeddings via a gated, element-wise modulation followed by layer normalization:

$$T_e = \text{LayerNorm}(T \odot (1 + \beta_T E_T)) \tag{3}$$

where $\odot$ denotes the Hadamard product and β_T is a learnable parameter controlling the infusion strength. This process yields emotion-aware text features T_e.

Visual Affective Cue Injection. Similarly, for the image I, we estimate its emotional content e_V. This can be achieved using a pre-trained visual emotion recognition model (e.g., trained on datasets like EMOTIC) or by leveraging CLIP's zero-shot capability with emotion-oriented textual descriptors (e.g., "a photo that evokes [emotion]"). The derived emotion score e_V is similarly projected to E_V and injected into the visual patch embeddings:

$$V_e = \text{LayerNorm}(V \odot (1 + \beta_V E_V)), \tag{4}$$

producing emotion-aware visual features V_e.

3.3 Cross-modal Inconsistency Modeling

The heart of CIAF is the **Relational Discrepancy Attention (RDA)** mechanism. Unlike standard cross-attention that seeks alignment, RDA is explicitly designed to *attend to and quantify discrepancies* between every text token and image patch pair across multiple complementary facets.

Discrepancy Score Computation. For a token t_i and a patch v_j, we compute three distinct similarity/consistency scores, each normalized to $[0,1]$: *(1) Semantic Similarity* S_{ij}^{sem}: Measures basic conceptual alignment using cosine similarity in the CLIP space. *(2) Emotional Consistency* S_{ij}^{emo}: Measures the alignment of emotional tones, defined as $1 - |e_T^{(i)} - e_V^{(j)}|$, where $e_T^{(i)}$ and $e_V^{(j)}$ are the emotion scores associated with the token and patch, respectively. *(3) Logic Consistency* S_{ij}^{log}: Estimates higher-order logical or event-based coherence. This can be derived from external knowledge (e.g., entity

co-occurrence graphs) or implicitly learned. For simplicity in initial modeling, we can set this as a learnable bias or estimate it via a lightweight network over $[t_i; v_j]$.

$$S_{ij}^{sem} = \frac{t_i \cdot v_j}{\|t_i\| \|v_j\|}, \quad S_{ij}^{emo} = 1 - |e_T^{(i)} - e_V^{(j)}|, \quad S_{ij}^{log} = \text{MLP}_{log}([t_i; v_j]). \tag{5}$$

Attention over Discrepancies. We then synthesize these scores into a holistic *relational discrepancy* score D_{ij}. Lower similarity implies higher discrepancy.

$$D_{ij} = \gamma_{sem}(1 - S_{ij}^{sem}) + \gamma_{emo}(1 - S_{ij}^{emo}) + \gamma_{log}(1 - S_{ij}^{log}), \tag{6}$$

where $\gamma_{sem}, \gamma_{emo}, \gamma_{log}$ are learnable parameters. These discrepancy scores are converted into an attention map $\mathcal{A}$ using a softmax operation scaled by a temperature factor τ:

$$\tilde{A}_{ij} = \exp(\tau D_{ij}), \quad \mathcal{A}_{ij} = \frac{\tilde{A}_{ij}}{\sum_{i'=1}^{n} \sum_{j'=1}^{m} \tilde{A}_{i'j'}}. \tag{7}$$

Crucially, this mechanism assigns higher attention weights to token-patch pairs that exhibit significant inconsistency across any of the measured facets.

Inconsistency Feature Aggregation. The final output of the RDA module is an aggregated inconsistency feature vector r. For each highly attended (inconsistent) pair (i, j), we construct a rich representation by concatenating their features, element-wise product (capturing interaction), and absolute difference (highlighting disparity). This representation is then processed by a small Multi-Layer Perceptron (MLP) $\psi(\cdot)$ and weighted by the attention $\mathcal{A}_{ij}$:

$$r = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{A}_{ij} \; \psi([t_i; v_j; t_i \odot v_j; |t_i - v_j|]), \tag{8}$$

where $\psi: \mathcal{R}^{4d} \to \mathcal{R}^{d_r}$ maps the concatenated vector to a lower-dimensional inconsistency representation.

3.4 Evidence-conditioned Consistency Reasoning

After obtaining the localized inconsistency representation from RDA, CIAF performs fact-checking through an evidence-conditioned consistency reasoning module. Instead of relying on a generic multimodal fusion vector, this module explicitly aggregates three types of evidence: the global textual representation, the global visual representation, and the discrepancy-aware representation produced by RDA. The final reasoning vector is constructed as:

$$x_f = [T_w; V_i; |T_w - V_i|; T_w \odot V_i; r], \tag{9}$$

where T_w and V_i denote the global textual and visual representations, r is the RDA-based inconsistency feature, and $[\cdot]$ denotes concatenation. This design forces the classifier to consider not only what each modality expresses independently, but also where and how the two modalities disagree.

The reasoning vector is then passed through a lightweight classifier:

$$P(y|x_f) = \text{Softmax}(\text{MLP}_{cls}(x_f)), \tag{10}$$

where y denotes the fact-checking label. Compared with coarse multimodal fusion, this module preserves the explicit discrepancy evidence produced by RDA and makes the final decision more closely tied to cross-modal consistency assessment.

3.5 Training Objective

To guide the learning of all components effectively, we employ a multi-task loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha_{con}\mathcal{L}_{con} + \alpha_{emo}\mathcal{L}_{emo} + \alpha_{reg}\mathcal{L}_{reg}. \tag{11}$$

Here, $\mathcal{L}_{CE}$ denotes the standard cross-entropy loss between the predicted probability $P(y|x_f)$ and the ground-truth label. $\mathcal{L}_{con}$ is an optional contrastive loss that pulls together global representations of text and image from the same sample while pushing apart those from different samples, thereby strengthening the shared embedding space. $\mathcal{L}_{emo}$ represents a consistency loss that penalizes large differences between the predicted emotional tones of the text and image modalities for a genuine news sample, based on the extracted e_T and e_V. Finally, $\mathcal{L}_{reg}$ is a regularization term (e.g., L2 weight decay) applied to model parameters to prevent overfitting. The coefficients α_* balance the contribution of each auxiliary objective, ensuring stable and convergent training.

4 Experiments

4.1 Dataset

To comprehensively evaluate our proposed CIAF framework, we conduct experiments on two widely used multimodal fact-checking benchmarks: PolitiFact and GossipCop. These datasets contain news articles paired with images and annotated with binary credibility labels (real or fake). PolitiFact consists of 1,056 verified political news items, whereas GossipCop is substantially larger, comprising 22,140 entertainment-related news articles. The diversity and scale difference between the two datasets enable a thorough assessment of the model's generalization across distinct domains.

Table 1 summarizes dataset statistics as well as the distributional characteristics (mean and standard deviation) of these sentiment features. The statistics reveal notable differences in emotional patterns between verified and false claims, providing empirical motivation for leveraging cross-modal discrepancy cues in our approach.

Table 1. Dataset statistics of PolitiFact and GossipCop.

Statistics	PolitiFact		GossipCop	
	fake	real	fake	real
Numberofnews	96	102	1877	4928
Averageofwords	444	3387	726	699

4.2 Baselines

To evaluate the performance of the proposed CIAF framework, we compare it with nine representative baseline models on the multimodal fact-checking benchmarks. These include: unimodal approaches, namely LDA-HAN [1] (which integrates LDA with a Hierarchical Attention Network) and T-BERT [2] (employing a cascaded triplet BERT architecture); multimodal approaches, including SAFE [30] (transforming images into textual descriptions), RIVF [23] (fusing VGG and BERT features via attention mechanisms), SpotFake [19] (leveraging VGG and BERT for feature extraction), and CAFE [5] (adopting fuzziness-aware feature aggregation); a standard fine-tuning approach based on RoBERTa [8]; and few-shot learning methods, namely P&A [27] (which jointly leverages pre-trained knowledge in PLMs and social-contextual topology, with the user engagement threshold adjusted to 50 when trained on external datasets due to the lack of user data in our own dataset) and SAMPLE [8] (a similarity-aware multimodal baseline designed for limited-data settings).

4.3 Implementation Details

We employ the AdamW optimizer with a cross-entropy loss function, an initial learning rate of 3×10^{-5}, and a weight decay of 1×10^{-3} over 20 epochs. Auxiliary affective cues are extracted from the text using TextBlob's sentiment analysis tool (SAT). As shown in Table 2 and Table 3, the model's performance is evaluated under both few-shot and data-rich settings. In the few-shot setting, both PolitiFact and GossipCop are trained with 50 samples per class (50-shot). Each experiment is repeated five times with different random seeds to ensure robustness. In the data-rich setting, the dataset is split into training, validation, and testing sets with a ratio of 8:1:1, and performance metrics are averaged after removing the highest and lowest values.

4.4 Main Results

Data-rich Performance Analysis. As shown in Table 2, under the conventional data-rich setting, our proposed CIAF framework achieves the best performance on both the PolitiFact and GossipCop benchmarks. Specifically, CIAF obtains an accuracy of 85.42% and an F1-score of 83.36% on PolitiFact, and an accuracy of 77.48% and an F1-score of 75.33% on GossipCop, outperforming all competing methods. These results directly validate the core idea introduced in the paper: explicitly modeling fine-

grained inconsistencies between text and images, together with emotional cues, can substantially enhance the effectiveness of multimodal fact-checking.

Table 2. Data-rich performance comparison across PolitiFact and GossipCop datasets.

Method	PolitiFact		GossipCop	
	Accuracy	F1	Accuracy	F1
LDA-HAN	74.36±0.36	70.24±0.42	60.18±0.51	54.15±0.39
T-BERT	75.12±0.32	71.08±0.38	74.23±0.45	61.42±0.47
SAFE	65.27±0.41	64.19±0.39	64.35±0.43	55.28±0.52
RIVF	45.18±0.55	43.27±0.58	61.42±0.49	51.36±0.44
Spotfake	73.45±0.38	71.33±0.36	73.28±0.41	43.24±0.61
CAFE	67.32±0.44	67.19±0.40	72.15±0.39	59.47±0.43
FTRB	84.27±0.29	79.36±0.34	69.42±0.46	63.28±0.45
P&A	66.48±0.47	63.35±0.49	71.39±0.42	73.42±0.38
SAMPLE	81.36±0.31	80.27±0.33	73.24±0.40	64.39±0.46
CIAF(Ours)	**85.42±0.27**	**83.36±0.30**	**77.48±0.37**	**75.33±0.35**

The notable advantage in F1-score further demonstrates that the Relational Discrepancy Attention (RDA) mechanism can accurately locate cross-modal contradictions in both semantic and auxiliary affective dimensions. Moreover, the robust performance gains on the noisier and more implicit GossipCop dataset highlight the value of incorporating auxiliary affective discrepancy signals to improve model robustness. Compared with strong baselines such as FTRB and SAMPLE, CIAF shows superior reasoning capability, especially on PolitiFact, which requires more rigorous fact verification, while avoiding the severe performance imbalance observed in methods like SpotFake on GossipCop.

Few-Shot Performance Analysis. Furthermore, as shown in Table 3, CIAF also exhibits excellent data efficiency and generalization ability under the highly challenging 50-shot few-shot setting. CIAF consistently achieves the highest accuracy and F1-score on both benchmarks, further confirming the central idea of the paper—modeling fine-grained cross-modal inconsistencies and emotional cues constitutes a highly data-efficient learning paradigm. When training samples are extremely limited, models can no longer rely on large-scale data to learn superficial statistical correlations; instead, they must capture the most discriminative contradictions. Through the RDA mechanism, CIAF dynamically focuses on cross-modal semantic and affective conflict regions, while the consistency reasoning module aggregates localized contradiction evidence, enabling the model to quickly learn patterns of "inconsistency".

Table 3. Few-shot performance comparison (50-shot) across two public fact-checking datasets.

Method	**PolitiFact(50-shot)**		**GossipCop(50-shot)**	
	Accuracy	F1	Accuracy	F1
LDA-HAN	63.24±0.52	61.38±0.55	45.27±0.68	49.42±0.62
T-BERT	69.35±0.48	69.28±0.46	61.43±0.53	52.36±0.59
SAFE	56.18±0.58	46.52±0.64	51.24±0.61	44.38±0.65
RIVF	49.42±0.62	43.27±0.68	31.35±0.72	29.48±0.74
Spotfake	73.56±0.45	73.42±0.43	49.36±0.63	48.52±0.64
CAFE	61.32±0.54	52.48±0.60	61.28±0.55	50.43±0.61
FTRB	81.46±0.41	77.38±0.44	64.52±0.52	52.47±0.58
P&A	61.48±0.56	57.36±0.58	59.42±0.57	58.49±0.56
SAMPLE	78.52±0.42	77.46±0.45	62.38±0.54	58.53±0.57
CIAF(Ours)	**80.68±0.39**	**80.54±0.40**	**66.47±0.49**	**63.52±0.53**

Comparisons with baseline models further highlight the advantages of CIAF. Text-only methods (e.g., T-BERT) and early multimodal fusion approaches (e.g., SAFE) suffer substantial degradation in the few-shot scenario. Even strong models designed for data-rich settings, such as FTRB and SAMPLE, are consistently outperformed by CIAF in the 50-shot setting. This indicates that CIAF's emphasis on fine-grained, multidimensional (semantic and affective) inconsistency detection is less prone to overfitting and generalizes better than global alignment or overly complex fusion strategies. Notably, on the noisier GossipCop dataset with more implicit contradictions, CIAF achieves the largest improvement in F1-score by leveraging auxiliary affective discrepancy estimation to introduce stable discriminative signals, further demonstrating the critical role of auxiliary affective cues in enhancing robustness in real-world scenarios.

4.5 Ablation Study

To assess the contribution of each core component in the proposed Cross-modal Inconsistency Aware Framework (CIAF), we conduct ablation studies on the PolitiFact and GossipCop benchmarks by progressively removing the emotional feature extraction module (-Emo), the Relational Discrepancy Attention mechanism (-RDA), and the evidence-conditioned consistency reasoning module (-Reasoning). As shown in Table 4, the full CIAF model consistently achieves the best performance on both datasets, and the removal of any component leads to a noticeable degradation, demonstrating the complementary roles of all modules in enhancing overall accuracy and generalization.

Table 4. Ablation study results on data-rich setting. "-Emo" removes the Emotional Element Extraction module, "-RDA" removes the Relational Discrepancy Attention mechanism, "-Reasoning" removes the evidence-conditioned consistency reasoning module.

Method	PolitiFact		GossipCop	
	Accuracy	F1	Accuracy	F1
CIAF(Ours)	**85.42±0.27**	**83.36±0.30**	**77.48±0.37**	**75.33±0.35**
-Emo	83.27±0.34	81.18±0.36	73.42±0.35	72.35±0.38
-RDA	82.48±0.38	80.32±0.40	72.39±0.41	70.47±0.43
-Reasoning	81.56±0.42	78.42±0.45	71.63±0.46	70.28±0.48

Effect of the Emotional Feature Extraction Module. Removing the emotional feature extraction module causes performance drops of varying degrees on both datasets, with a more pronounced decline on GossipCop. This finding suggests that emotional cues serve as important auxiliary signals for detecting inconsistencies in cross-modal emotional expression, especially in entertainment-related misinformation where sentiment exaggeration is common.

Importance of the Relational Discrepancy Attention Mechanism. Eliminating the RDA module further reduces performance, particularly on PolitiFact. This highlights the significance of fine-grained relational modeling in locating subtle semantic and emotional contradictions. Without RDA, the model loses its ability to explicitly capture local cross-modal inconsistencies, thereby weakening its ability to detect complex multimodal conflicts.

Generalization Benefits of Evidence-conditioned Consistency Reasoning. Removing the evidence-conditioned consistency reasoning module leads to the most significant performance degradation. This module provides targeted aggregation of localized discrepancy evidence and improves generalization, especially in low-resource settings. By explicitly organizing token-patch contradiction signals before classification, it enhances robustness and improves consistency on unseen samples. Overall, the varying impact of each module across datasets reflects the domain-specific characteristics of misinformation. Emotional cues contribute more prominently to performance on GossipCop, while relational inconsistency modeling and structured reasoning play larger roles on the politically oriented PolitiFact dataset. These findings demonstrate the adaptability and effectiveness of CIAF's modular design across diverse misinformation scenarios.

5 Conclusion

This paper introduces a Cross-modal Inconsistency Aware Framework (CIAF) for multimodal fact-checking on social media. CIAF explicitly models fine-grained inconsistencies between text and images, addressing limitations of existing methods that rely mainly on global semantic alignment or simple feature fusion. By integrating a Relational Discrepancy Attention mechanism to locate inconsistent token-patch interactions and using evidence-conditioned consistency reasoning to structure verification decisions, CIAF enhances both accuracy and interpretability. Experiments on two challenging benchmarks demonstrate its clear advantages. Nonetheless, its generalization to more complex modalities such as video and its robustness under extreme few-shot conditions remain open challenges. Future work may extend cross-modal inconsistency modeling to temporal and conversational multimodal settings and leverage external knowledge for improved semantic and logical contradiction detection.

References

1. Comparing topic-aware neural networks for bias detection of news. In: Proceedings of 24th European Conference on Artificial Intelligence (ECAI 2020). vol. 325, pp. 2054-2061. International Joint Conferences on Artificial Intelligence (IJCAI) (2020)
2. Bhatt, S., Goenka, N., Kalra, S., Sharma, Y.: Fake news detection: Experiments and approaches beyond linguistic features. In: Data Management, Analytics and Innovation: Proceedings of ICDMAI 2021, Volume 2, pp. 113-128. Springer (2021)
3. Cao, M., Hu, P., Wang, Y., Gu, J., Tang, H., Zhao, H., Dong, J., Yu, W., Zhang, G., Reid, I., et al.: Video simpleqa: Towards factuality evaluation in large video language models. CoRR (2025)
4. Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., Wang, W.: mclip: Multilingual clip via cross-lingual transfer. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13028-13043 (2023)
5. Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., Shang, L.: Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM web conference 2022. pp. 2897-2905 (2022)
6. Choi, E.C., Ferrara, E.: Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In: Companion Proceedings of the ACM Web Conference 2024. pp. 1441-1449 (2024)
7. Guo, Z., Schlichtkrull, M., Vlachos, A.: A survey on automated fact-checking. Transactions of the association for computational linguistics **10**, 178-206 (2022)
8. Jiang, Y., Yu, X., Wang, Y., Xu, X., Song, X., Maynard, D.: Similarity-aware multimodal prompt learning for fake news detection. Information Sciences **647**, 119446 (2023)
9. La, T.V., Tran, Q.T., Tran, T.P., Tran, A.D., Dang-Nguyen, D.T., Dao, M.S.: Multimodal cheapfakes detection by utilizing image captioning for global context. In: Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval. pp. 9-16 (2022)
10. Li, J., Bin, Y., Zou, J., Wei, J., Wang, G., Yang, Y.: Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia. pp. 1-7 (2023)

11. Papadopoulos, S.I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C.: Red-dot: Multimodal fact-checking via relevant evidence detection. IEEE Transactions on Computational Social Systems (2025)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748-8763. PmLR (2021)
13. Rayar, F., Delalandre, M., Le, V.H.: A large-scale tv video and metadata database for french political content analysis and fact-checking. In: Proceedings of the 19th International Conference on Content-based Multimedia Indexing. pp. 181-185 (2022)
14. Saakyan, A., Chakrabarty, T., Muresan, S.: Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2116-2129 (2021)
15. Shang, L., Kou, Z., Zhang, Y., Wang, D.: A duo-generative approach to explainable multimodal covid-19 misinformation detection. In: Proceedings of the ACM Web Conference 2022. pp. 3623-3631 (2022)
16. Shao, R., Wu, T., Wu, J., Nie, L., Liu, Z.: Detecting and grounding multi-modal media manipulation and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5556-5574 (2024)
17. Shao, Y., Sun, J., Zhang, T., Jiang, Y., Ma, J., Li, J.: Fake news detection based on multimodal classifier ensemble. In: Proceedings of the 1st International Workshop on Multimedia Al against Disinformation. pp. 78-86 (2022)
18. Singhal, S., Dhawan, M., Shah, R.R., Kumaraguru, P.: Inter-modality discordance for multimodal fake news detection. In: Proceedings of the 3rd ACM International Conference on Multimedia in Asia. pp. 1-7 (2021)
19. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multimodal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). pp. 39-47. IEEE (2019)
20. Tahmasebi, S., Hakimov, S., Ewerth, R., Müller-Budack, E.: Improving generalization for multimodal fake news detection. In: Proceedings of the 2023 ACM international conference on multimedia retrieval. pp. 581-585 (2023)
21. Tang, C.W., Chen, T.C., Nguyen, K., Mehrab, K.S., Ishmam, A.M., Thomas, C.: M3d: Multimodal multidocument fine-grained inconsistency detection. In: EMNLP (2024)
22. Tang, L., Laban, P., Durrett, G.: Minicheck: Efficient fact-checking of llms on grounding documents. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 8818-8847 (2024)
23. Van Tuan, M., Budo, K., Tuan, C.A., Takeshita, T.: Analysis of unidirectional isolated yy connection three-phase dc-dc converter. In: 2021 IEEE International Future Energy Electronics Conference (IFEEC). pp. 1-6. IEEE (2021)
24. Wang, B., Li, X., Li, C., Wang, S., Gao, W.: Escaping the neutralization effect of modality features fusion in multimodal fake news detection. Information Fusion 111, 102500 (2024)
25. Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., Wang, S.: Cross-modal contrastive learning for multimodal fake news detection. In: Proceedings of the 31st ACM international conference on multimedia. pp. 5696-5704 (2023)
26. Wang, S., Lin, H., Luo, Z., Ye, Z., Chen, G., Ma, J.: Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. In: ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models

27. Wu, J., Li, S., Deng, A., Xiong, M., Hooi, B.: Prompt-and-align: prompt-based social alignment for few-shot fake news detection. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 2726-2736 (2023)
28. Ying, Q., Hu, X., Zhou, Y., Qian, Z., Zeng, D., Ge, S.: Bootstrapping multi-view representations for fake news detection. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 37, pp. 5384-5392 (2023)
29. Zeng, Z., Mao, W.: A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval. arXiv preprint arXiv:2201.02772 (2022)
30. Zhou, X., Wu, J., Zafarani, R.: Similarity-aware multi-modal fake news detection. In: Pacific-Asia Conference on knowledge discovery and data mining. pp. 354-367. Springer (2020)
31. Zhu, Z., Zhang, H., Wu, G., Lyu, S., Wu, B.: Hmgie: Hierarchical and multi-grained inconsistency evaluation for vision-language data cleansing. arXiv preprint arXiv:2412.05685 (2024)