ERCA: A Fake News Detection Framework based on Multi-Agent-Based Elemental Reconstruction and Consistency Analysis

Anonymous ACL submission

Abstract

Multimodal fake news detection has become a research focus in the context of the rapid spread of misleading content. However, existing methods face two key challenges: (1) the methods usually relying on the direct cross-modal feature alignment of the model, which cannot fully capture the semantic nuances between text and images; (2) The public datasets cannot focus on scenarios for consistency analysis and lack hierarchical difficulty labels. To address the problems, we classify the text-image semantic inconsistency based on its complexity into three difficulty levels and design different methods to generate approximately 50,000 multimodal fake news samples. Meanwhile, we propose a Multi-Agent-Based Elemental Reconstruction and Consistency Analysis Framework(ERCA), which transforms the visual semantic elements in the image into the text modal based on the original text constraints, thereby minimizing the modal differences and information loss during the reasoning stage. We also design a new interaction mode for the multi-agent interaction stage, which allows Large Visual Language Model to perform fine-grained verification at the element level instead of exchanging the entire image information. A large number of experiments show that ERCA outperforms the existing benchmark methods on our dataset, especially in medium and difficult datasets.

1 Introduction

With the explosive growth of multimodal content on social media and digital platforms, the spread of fake news has posed severe threats to social stability, public trust, and information ecology(Khan et al., 2021; Tian et al., 2025). Multimodal fake news detection aims to identify such deceptive content by analyzing the correlation and consistency between textual and visual modalities (Nasser et al., 2025; Zhou and Zafarani, 2020). Meanwhile, multi-agent frameworks, which leverage the collaborative capabilities of multiple intelligent models to tackle complex tasks, have emerged as a promising paradigm in artificial intelligence, offering enhanced robustness and granularity compared to single-agent systems (Shi and Yan, 2021; Tran et al., 2025; Li et al., 2024a). The integration of multiagent systems and multimodal fake news detection has become a key direction to address the limitations of traditional detection methods (Liu et al., 2025).

Although significant progress has been made in the field of multimodal fake news detection recently, the existing methods still have some key challenges that need to be addressed. These challenges hinder the practicality and performance of the methods. First, semantic inconsistency capture remains insufficient. Most traditional methods rely on direct cross-modal feature alignment (He et al., 2025; Shen et al., 2024)or shallow consistency checks (Xue et al., 2021; Jin et al., 2024). Second, in the public datasets, false news is classified as negative samples for various reasons, including factual errors, consistency errors, etc (Nan et al., 2021). This results in the dataset being unable to effectively verify the performance of the method in multimodal consistency analysis. Moreover, most existing multimodal fake news datasets, such as FakenewsNet (Shu et al., 2018) and FMNV (Wang et al., 2025), only label the samples as "true" or "false", without considering the complexity of the inconsistency between text and images, lacking hierarchical difficulty labels, and thus unable to evaluate the robustness of the detection model in different difficulty scenarios. Additionally, most current detection methods based on LVLM when generating image descriptions or analyzing visual content lack the awareness of decomposing complex images into individual elements. Some methods (Lv et al., 2025; Xie et al., 2025) adopt the heterogeneous graph approach for modality mapping, but the methods are complex, heavily dependent

on model structure design, and lack generalization. In the context of multi-agent frameworks for fake news detection, existing studies (Li et al., 2024b; Verma et al., 2025) primarily focus on multi-Large Language Models(LLM) collaboration for textual misinformation, with few extending to multimodal scenarios.

To address the above challenges, this paper proposes a multi-agent-based framework for elemental reconstruction and consistency analysis(ERCA). Specifically, the framework abandons traditional cross-modal feature alignment and instead adopts a textual reconstruction-driven consistency analysis approach. Unlike feature alignment methods that operate in high-dimensional latent spaces with limited interpretability and vulnerability to irrelevant noise, our element reconstruction decomposes visual information into discrete semantic units, forcing the model to focus on fine-grained consistency. This design not only enhances sensitivity to subtle tampering as confirmed by experiments, but also offers better interpretability and lower training costs compared to end-to-end alignment. The framework uses LLM to extract key event elements from the original news text as constraints. Then, through the Large Vision Language Models(LVLMs) multi-agent system, it performs the verification, extraction and reconstruction of visual information based on the original text constraints, converting it into text form. The consistency between the original news text and the reconstructed image text is then evaluated by a fine-tuned LLM, enabling the capture of subtle event-level discrepancies. Furthermore, to solve the lack of hierarchical datasets, we design a dedicated method to construct a large-scale hierarchical multimodal fake news dataset based on publicly available Twitter data. The main contributions of our work can be summarized as follows:

- We designed different methods based on the complexity of semantic inconsistency between text and images to generate approximately 50,000 multimodal fake news samples. The corpus focused on the complexity of inconsistency, which can strictly evaluate the consistency analysis ability and robustness of our method.

- Unlike traditional cross-modal alignment methods, our framework leverages the synergy between LLMs and LVLMs multi-agent to convert visual semantic elements into textual form under the constraint of original news text. This design minimizes modal differences and information loss during reasoning.

- We have proposed a new interaction mechanism. In the multi-agent module, the visual language model will exchange and verify individual visual elements rather than the entire image information. This element-level fine-grained interaction ensures that each component of the image can be independently verified.

2 Related Works

2.1 Multimodal Fake News Detection

Early multimodal fake news detection methods primarily fuse features independently extracted from text and visual modalities. Representative works include EANN, which introduces an event discriminator to learn event-invariant cross-modal features (Wang et al., 2018), and MVAE, which employs a bimodal variational autoencoder to learn shared text-image representations (Khattar et al., 2019). Subsequent methods have focused on explicitly modeling cross-modal relationships: CAFÉ adaptively aggregates unimodal and cross-modal features based on sample-level ambiguity (Chen et al., 2022), and SR-CIBN dynamically adjusts the balance between consistent and inconsistent features according to image-text matching degree (Yu et al., 2025). Despite this progress, these methods operate through global feature alignment in high-dimensional latent spaces, offering limited interpretability and struggling to capture fine-grained element-level inconsistencies that characterize subtle fake news.

More recent work has pushed toward fine-grained and uncertainty-aware detection paradigms. Li et al. proposed UDRL, an uncertainty-aware disentangled representation learning framework that separates modality-specific and cross-modal features while explicitly modeling prediction confidence, thereby improving robustness against ambiguous news samples (Li et al., 2025). From a data-centric perspective, the AMG benchmark introduces attribution-driven multi-granularity annotations that categorize fake news into distinct falsification patterns, moving beyond binary labels to capture the diverse ways in which multimodal news can be deceptive (Guo et al., 2025). Our

work shares this spirit of fine-grained analysis but differs fundamentally by decomposing visual content into discrete semantic elements and verifying them through multi-agent collaboration under textual constraints.

2.2 Large Vision-Language Models and Multi-Agent Systems

The emergence of LVLMs has advanced fine-grained visual understanding. Recent works such as DeepScan boost LVLM performance on detailed visual tasks through hierarchical scanning and evidence-enhanced reasoning (Li et al., 2026), while GranViT achieves state-of-the-art fine-grained recognition via region-level autoregressive training (Zheng et al., 2025). In the specific context of misinformation, ContextGuard-LVLM proposes a multi-stage contextual reasoning framework built upon LVLMs that verifies news authenticity through fine-grained cross-modal consistency checking at the entity and relation levels (Ma et al., 2025). These methods demonstrate the growing capability of LVLMs to go beyond holistic image understanding toward structured, element-aware visual analysis.

In parallel, multi-agent collaboration has been applied to misinformation detection. RAMA leverages multiple multimodal LLMs with complementary prompts for fact-checking (Yang et al., 2025), and AgentFact emulates the human verification workflow through five steps include strategy planning, evidence retrieval, visual analysis, reasoning, and evidence-grounded explanation generation combined by retrieval-reasoning (Xu et al., 2025a). Multi-agent debate protocols have also proven effective for reducing hallucinations in LLMs through structured peer review (Huang et al., 2026). Extending this direction, FakeAgent introduces a Delphi-inspired multi-agent reasoning framework that integrates multimodal understanding with external evidence retrieval for attribution-grounded misinformation analysis in micro-videos (Zeng et al., 2026). However, existing multi-agent frameworks for misinformation detection primarily focus on textual claims or holistic image-text matching, with limited work addressing fine-grained element-level cross-verification between modalities—a gap that our ERCA framework directly targets.

2.3 Datasets and Element-Level Analysis

High-quality datasets are crucial for advancing multimodal fake news detection. Widely used benchmarks such as Twitter, Weibo, and FakeNewsNet (Shu et al., 2018) provide real-world multimodal samples with authenticity labels. However, they classify false news as negative samples for mixed reasons and only offer binary labels without differentiating the complexity of text-image inconsistency. Recent works have recognized these limitations: NewsCLIPpings automatically generates out-of-context image-text pairs for fine-grained evaluation (Luo et al., 2021), and MAGNet constructs heterogeneous graphs with modality-specific edges to model fine-grained cross-modal features (Wang et al., 2026). FND-EKA further introduces hierarchical conditional detection with external knowledge (Huang et al., 2025). Nevertheless, these graph-based methods are complex and lack generalization, and no existing public dataset provides explicit hierarchical difficulty labels.

Several recent dataset efforts have made progress toward multi-level evaluation. The MMM-Fact benchmark introduces retrieval difficulty levels for multimodal fact-checking by varying the complexity of evidence aggregation required for veracity prediction (Xu et al., 2025b). The M4FC dataset spans six multimodal fact-checking tasks across multiple languages and cultural contexts, highlighting the need for diverse evaluation protocols (Geng et al., 2025). Most notably, the AMG benchmark explicitly annotates each fake sample with its falsification type and granularity level, providing the first systematic effort toward attribution-based difficulty stratification (Guo et al., 2025). However, AMG's difficulty taxonomy is defined by falsification category rather than by the complexity of semantic inconsistency between text and images. Our work fills this remaining gap by constructing a large-scale dataset with three explicitly defined difficulty levels—simple, medium, and hard—based on the degree and subtlety of text-image semantic mismatch, specifically designed to evaluate fine-grained consistency analysis capabilities.

3 Method

3.1 Problem Definition

The multimodal fake news detection task can be formalized as a binary classification problem, where multimodal news samples are mapped to authenticity labels (Jing et al., 2023). The input is a

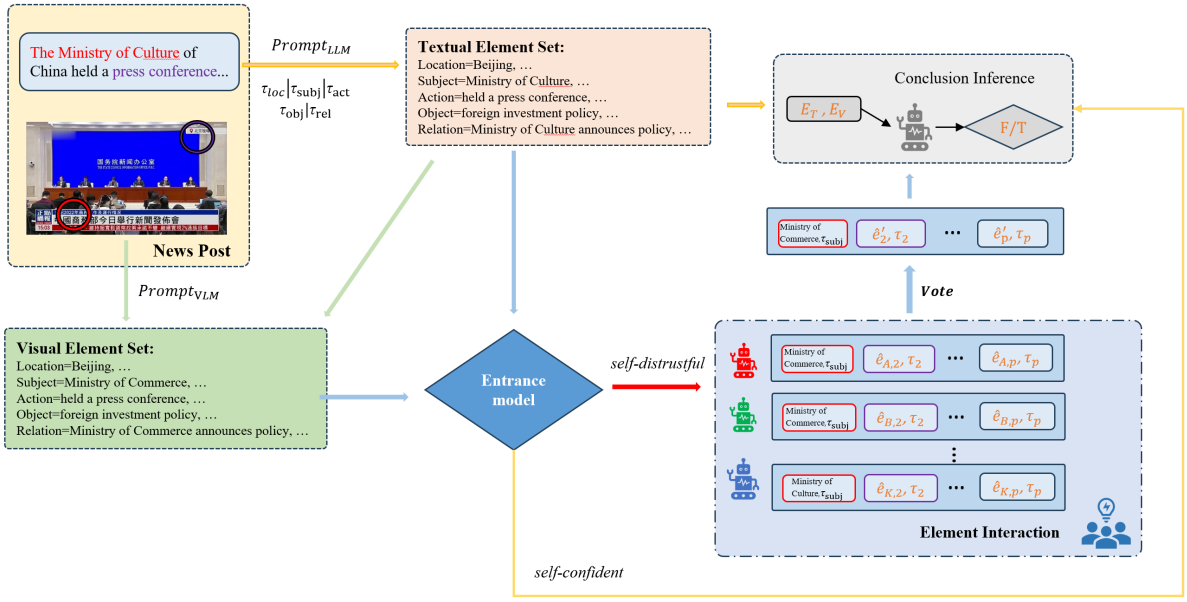

Figure 1: Overview of ERCA. The proposed framework achieves multimodal consistency analysis through element extraction → fine-grained verification → conclusion inference. The LLM extracts the core elements of the event from the original news text, setting a semantic benchmark for subsequent visual analysis; the multi-agent composed of prompt-tuned LVLMs independently extract image elements, then engage in element-level cross-validation through majority voting to ensure accuracy and generate unified reconstructed text. The fine-tuned LLM outputs the result based on the consistency calculation of the text elements and the reconstructed text, minimizing the modality difference and capturing fine details.

news post to be verified, which can be denoted as $S = (T, V)$, where T represents the original textual content of the news, V represents the visual content of the news (an image).The goal of multi-modal fake news detection is to learn a classifier $f : (T, V) \rightarrow Y$, where $Y = \{0, 1\}$ is the binary label (0 = fake news, 1 = real news).

The overall framework of ERCA is illustrated in Fig.1. In the subsequent subsections, we discuss each module in detail.

3.2 Element Extraction

The element extraction module uses a LLM fine-tuned with instructions to extract the textual event element set E_T from the original text T. This module ensures that subsequent visual analysis is anchored to the core semantics of the news, avoiding "off-topic" extraction of visual elements. The LLM is instruction-tuned on 200 manually labeled samples solely to stabilize the output format; this step does not affect the generality of our method.

The event key elements extracted from T are defined as a set $E_T = \{e_1, e_2, ..., e_p\}$, where p is the number of distinct elements. Each element e_i belongs to a predefined semantic type τ_i, and it is represented as a tuple (v_i, τ_i), where v_i is the element value and τ_i is its semantic type. We define 5 core semantic types for news events, covering the most critical dimensions of inconsistency: location of the event τ_{loc}; subject of the event τ_{subj}; action of the event τ_{act}; key objects in the event τ_{obj}; relation of entity in the event τ_{rel}.

For the visual content V, the visual semantic elements extracted by LVLMs are denoted as $E_V = \{\hat{e}_1, \hat{e}_2, ..., \hat{e}_p\}$. It is important to note that in fake news samples, the visual elements E_V may be missing, incorrect, or contain extra elements compared to E_T, which naturally introduces asymmetry. Our framework is designed to handle such mismatches robustly, as demonstrated in the experiments.

We use a prompt-tuned LLM with a structured $\text{Prompt}_{\text{LLM}}$ to enforce consistent element extraction. The $\text{Prompt}_{\text{LLM}}$ makes model give the news text T, extract its key event elements into 5 categories: Location, Subject, Action, Object and Relation. For each category, output 1-3 most relevant elements. Each element description should not exceed 7 words. If a category is unavailable, output 'N/A'."

The LLM processes $\text{Prompt}_{\text{LLM}}$ and outputs a structured response, which is parsed into the textual

element set $E_T = \{e_1, e_2, ..., e_k\}$.

The model is fine-tuned based on the LORA method, and the loss function is as follows, where $\mathbf{z}_{i,k}$ is the output by the LoRA-fine-tuned LLM for the k-th element type of the i-th sample and $\mathbf{y}_{i,k}$ is the ground-truth label for the k-th element type of the i-th sample:

$$\mathcal{L}_{\text{LLM-elem}} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \text{CE}\left(\text{softmax}\left(\mathbf{z}_{i,k}\right), \mathbf{y}_{i,k}\right)}{N \cdot K} \tag{1}$$

3.3 Fine-grained verification

The multi-agent verification module is activated when the LVLMs entry model determines that the sample requires further validation (Section 3.3.1 defines the entry model's decision logic). This module uses K LVLMs (denoted as $A_1, A_2, ..., A_K$, $K \geq 3$ to ensure diversity) to extract and cross-verify visual elements, generating a reliable visual element set E'_V.

3.3.1 Entrance Model

This module is designed to accelerate the process of false information detection, avoiding the waste of resources in simple scenarios. The Entrance model is a prompt-tuned LVLM, denoted as A_0. It takes two inputs: the original text T and the textual element set E_T.

The prompt for A_0 is Prompt_{A_0} , which makes the Entrance model give the news event elements E_T , for each type e_i in E_T, extract elements of the same semantic types from image V. If no element matches, output 'N/A'. Then, Judge if multi-agent verification is needed. A_0 outputs $E_{V,K}$ with a binary trigger signal $\delta \in \{0, 1\}$, where: $\delta = 0$: No multi-agent verification needed , direct inference via LLM; $\delta = 1$: Multi-agent verification needed, inference via LVLMs-agents.

3.3.2 Multi-Agent Architecture and Element Interaction

Rationale for Multi-Agent Design. The adoption of a multi-agent architecture is motivated by four key considerations. **(1) Enhanced fine-grained verification and adversarial robustness:** Element-level verification demands high precision; a single LVLM is prone to stochastic errors or prompt sensitivity (Schmalfuss et al., 2025). Multiple independently initialized models voting or discussing can filter out random noise and individual biases, raising confidence in each element's presence judgment. Moreover, fake news creators may employ adversarial attacks (e.g., subtly altering key objects via AIGC) targeting a specific model (Liu et al., 2026). It is considerably harder for an attacker to simultaneously deceive three LVLMs with different architectures and initializations. Even if one agent is misled, the correct outputs from others can be preserved through cross-verification. **(2) Unsupervised error correction via consensus:** Without external knowledge bases, multi-agent interaction enables consensus-based self-supervision. If a majority of agents deem an element absent, the element can be vetoed even if extracted by one agent. This mechanism counters hallucination—model-generated outputs not grounded in visual facts—such as a case where Llama3.2-Vision misidentified a white car as an ambulance in our preliminary experiments. **(3) Mitigation of inherent data bias and attention preferences:** Every LVLM carries biases from its training data and distinct attention preferences. For instance, Llama3.2-Vision may attend more to "actions", while Qwen2.5-VL focuses more on "objects". Multiple agents with complementary perspectives achieve more balanced element coverage (Hong and Quan, 2026). **(4) Improved interpretability:** Unlike end-to-end black-box single models, the multi-agent system records each agent's extraction results and the discussion process, producing an intermediate evidence chain: which agent extracted which element, which elements were vetoed during cross-verification, and which semantic category (τ_{loc}, τ_{act}, etc.) caused the final inconsistency. This provides interpretable intermediate states valuable for subsequent human auditing and model debugging (Du et al., 2024).

The multi-agent system consists of K(in this paper, we set K = 3) homogeneous LVLMs with identical capabilities but independent initializations.

In the first step, each agent A_k ($k = \{1, 2, 3\}$) takes the image V and textual element set E_T as inputs, and extracts a visual element candidate set $E_{V,k}$ via the $\text{Prompt}_{\text{LVLM}}$, which makes model give the news event elements E_T , for each type e_i in E_T, extract elements of the same semantic types from image V. If no element matches, output 'N/A'."

Each $E_{V,k} = \{\hat{e}_{k,1}, \hat{e}_{k,2}, ..., \hat{e}_{k,p}\}$, where p is the number of semantic elements in E_T, and $\hat{e}_{k,j} = (v_{k,j}, \tau_j)$.

Then, for each semantic element position j, the agents engage in a structured discussion based on

a modified voting mechanism to avoid multi-round inefficiency. Each agent A_k checks whether its extracted element $\hat{e}_{k,j}$ is truly present in V and produces a binary validity vote. The final element $\hat{e}'_j$ is determined by majority rule: if at least two agents agree on an element value, it is accepted; otherwise, the agents perform a single round of cross-verification by comparing their observed details, and the element supported by the most evidence is selected. If no consensus is reached, the element is marked as "N/A". This process yields the verified visual element set $E'_V = \{\hat{e}'_1, \hat{e}'_2, ..., \hat{e}'_p\}$.

3.4 Conclusion Inference

The $\text{LLM}_{\text{finetune}}$, denoted as f, is trained on our hierarchical dataset(defined in Section 4.1) to learn the mapping $(E_T, E'_V) \rightarrow Y$. The classifier f makes decisions based on the consistency between textual event elements E_T and visual semantic elements E'_V. We define the consistency set $S(E_T, E'_V) = \{y_1, y_2, ..., y_k\}, y_k = \{0, 1\}$, where y_k represents the consistency label of the k-th element. The LLM outputs Y based on the classification rule as follows:

$$Y = \begin{cases} 0 & \text{if } \exists y \in S(E_T, E_V) = 0, \\ 1 & \text{if } \forall y \in S(E_T, E_V) = 1, \end{cases} \quad (2)$$

4 Experiment

4.1 Dataset Construction Pipeline

We mainly use publicly available Twitter and Weibo datasets as reliable sources to collect real multimodal news samples to ensure the authenticity of the basic content. Each sample contains a complete news text T_{real} and the corresponding real image V_{real}. We eliminate low-quality samples, such as unreadable text and blurry images. For each real sample $(T_{\text{real}}, V_{\text{real}})$, we generate a fake sample by introducing controllable inconsistencies between the text and the modified image V_{fake}. The injection strategy varies according to the difficulty level to ensure the presence of distinguishable complexity.

For the simple level, we insert obvious, non-overlapping inconsistencies. We randomly replace V_{real} with images unrelated to the event. The semantic gap between T_{real} and V_{fake} is significant, making the inconsistencies easy to detect. For the medium difficulty level, we introduce inconsistencies at the element level. We modify 1-2 key event elements in V_{real}. The remaining elements remain consistent, which requires the model to capture subtle differences. For the hard level, we first summarize the linguistic and visual patterns that cause failures in baseline models on a held-out validation set. These patterns include multi-step reasoning requirements and deep context understanding. We then manually screen additional real samples and introduce modifications based on these patterns, ensuring the resulting hard samples are challenging yet not overfitted to specific models. The modified elements are very subtle visually and require in-depth semantic analysis.

By applying the aforementioned strategy, we generated a total of about 50,000 fake samples. The dataset contains 25,000 simple-level, 20,000 medium-level, and 4,000 hard-level samples. A balanced number of real samples is included for each level, ensuring class balance.

4.2 Experiment Setup

To ensure fairness and reproducibility, we detail the model training parameters, hardware configuration, baseline methods and evaluation metrics below.

Taking into account both performance and computational cost, we chose Qwen2.5-VL (Bai et al., 2025), QVQ-72B-Preview and Llama3.2-Vision (Lee et al., 2025) as the basic models for our LVLM multi-agent system. To ensure semantic consistency, we used QVQ-72B-Preview as the entrance model. Furthermore, the reasoning of the text modality utilizes llama3-70B (Dubey et al., 2024). All the models are trained and evaluated on our dataset with a dataset split of 6:2:2. We implemented our model using PyTorch and ran the experiments on NVIDIA A100 (80GB VRAM). The model is trained using the Adam optimizer with a learning rate schedule: an initial rate of 5e-2 is used for the first 2 epochs to allow the model to quickly adapt to the task, after which it is reduced to 1e-5 for the remaining epochs. The total number of training epochs is 30. All baselines are fine-tuned on our dataset training set under the same hardware conditions to ensure fair comparison.

We use three standard metrics to evaluate detection performance: Precision (Pre.), Recall (Rec.), and F1-Score (F1). We report metrics for the entire test set (Overall) and each difficulty level separately to assess model robustness.

4.3 Result

As shown in Table 1, our ERCA framework consistently outperforms all baseline methods and in-

	simple			medium			hard		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
EANN-KDD18 (Wang et al., 2018)	0.831	0.816	0.823	0.638	0.621	0.629	0.371	0.290	0.326
MVAE (Khattar et al., 2019)	0.857	0.845	0.851	0.693	0.677	0.685	0.414	0.362	0.386
CAFÉ (Chen et al., 2022)	0.883	0.862	0.872	0.673	0.681	0.677	0.399	0.328	0.360
Event-Radar (Ma et al., 2024)	0.954	0.943	0.948	0.754	0.726	0.740	0.573	0.549	0.561
Llama3.2-Vision	0.926	0.914	0.920	0.685	0.662	0.673	0.519	0.503	0.511
Qwen2.5-VL	0.933	0.917	0.925	0.704	0.699	0.701	0.547	0.531	0.539
QVQ	0.943	0.925	0.934	0.729	0.707	0.718	0.572	0.538	0.554
ERCA	**0.951**	**0.944**	**0.947**	**0.766**	**0.730**	**0.748**	**0.604**	**0.562**	**0.582**

Table 1: Performance comparison on our hierarchical dataset. Best results are in **bold**. (P: Precision, R: Recall, F: F1 Score)

dividual LVLMs across all difficulty levels. These results are the median values obtained through multiple repeated experiments. The single LVLM models (Llama3.2-Vision, Qwen2.5-VL, QVQ) achieve moderate performance, with QVQ being the strongest among them, but still significantly lower than ERCA, especially on hard samples. This gap validates the advantage of multi-agent collaboration: by integrating complementary attention preferences and cross-verifying elements, ERCA mitigates individual model biases and hallucinations, leading to more robust fine-grained detection.

In simple-level samples, all methods achieve high F1 scores; the advantage of ERCA is marginal over Event-Radar. However, when moving to medium and hard levels, performance drops sharply for all single models and baselines, while ERCA maintains the highest F1 in both settings. Notably, the hard-level F1 of ERCA exceeds the best baseline Event-Radar by 2.1 percentage points, and outperforms the best single LVLM (QVQ) by 2.8 percentage points. This demonstrates the framework's superior ability to capture subtle element-level inconsistencies that require multi-step reasoning and contextual understanding.

4.4 Generalization and Efficiency Analysis

To further assess generalization, we evaluate all methods on three public datasets (Twitter, Weibo, Fakenewsnet) and on the entire self-constructed dataset (all difficulty levels combined). Table 2 reports F1, Accuracy, and inference delay (the ratio to the standard, measured on the same A100 hardware). ERCA achieves the best performance across all datasets, with an F1 of 0.907 on Twitter, 0.933 on Weibo, and 0.842 on Fakenewsnet, significantly surpassing the strongest baseline Event-Radar. This consistent improvement demonstrates that our element-reconstruction-based approach is not tailored to a specific data distribution but generalizes well to diverse multimodal news sources.

We also report the variant ERCA w/o entrance (i.e., removing the entrance model and routing all samples through multi-agent verification). While this variant yields slightly higher F1 scores, its inference delay is 1.84, nearly double that of the full ERCA. The full ERCA benefits from the entrance model, which filters out a large portion of easy samples, triggering the costly multi-agent interaction only when necessary. This design achieves an excellent trade-off: ERCA retains 98% of the w/o entrance model's peak performance while reducing latency by 45.7%. Compared to single LVLMs like QVQ, ERCA adds only 0.08 delay but delivers substantially higher accuracy, especially on hard cases. Thus, the entrance model is critical for practical deployment, enabling high throughput without sacrificing detection reliability.

4.5 Ablation Study

We conducted an ablation study on our hierarchical dataset to evaluate the impact of key components: multi-agent interaction, the entrance model, model pre-training, and image-to-text reconstruction. As shown in Table 3, omitting any of these modules led to a lower F1-score across all difficulty levels, confirming the necessity of each component.

Omitting the text-res module caused the most severe performance degradation across all difficulty levels. This underscores the irreplaceable role of converting visual semantics into textual form: without this module, the framework relies solely on coarse feature alignment, failing to capture subtle element-level inconsistencies. Omitting the pre-train module also resulted in a substantial drop, particularly on hard samples, indicating that pre-trained capabilities are essential for accurate element extraction and fine-grained detail recognition in complex scenarios.

	Twitter		Weibo		Fakenewsnet		self-dataset-all		Inference
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	delay
EANN-KDD18	0.796	0.811	0.822	0.837	0.745	0.758	0.701	0.725	0.59
MVAE	0.821	0.836	0.847	0.862	0.772	0.783	0.743	0.748	0.65
CAFÉ	0.844	0.859	0.871	0.885	0.793	0.806	0.748	0.771	0.68
Event-Radar	0.873	0.898	0.915	0.924	0.822	0.835	0.829	0.837	0.72
Llama3.2-Vision	0.791	0.847	0.857	0.873	0.735	0.791	0.783	0.791	0.76
Qwen2.5-VL	0.845	0.849	0.895	0.910	0.818	0.833	0.800	0.809	0.81
QVQ	0.865	0.876	0.913	0.919	0.834	0.849	0.805	0.812	0.92
ERCA	0.907	0.914	0.933	0.945	0.842	0.853	0.834	0.846	1.00
ERCA w/o entrance	0.910	0.917	0.939	0.950	0.847	0.861	0.839	0.851	1.84

Table 2: Cross-dataset performance (F1 and Accuracy) and inference delay (the ratio to the standard).

	simple			medium			hard		
	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
w/o agent	0.943	0.925	0.934	0.729	0.707	0.718	0.572	0.538	0.554
ERCA w/o entrance	0.945	0.926	0.935	0.732	0.712	0.722	0.576	0.545	0.560
w/o pre-train	0.902	0.879	0.890	0.684	0.661	0.672	0.526	0.472	0.498
w/o text-res	0.860	0.827	0.843	0.649	0.637	0.643	0.471	0.424	0.446
our	**0.951**	**0.944**	**0.947**	**0.766**	**0.730**	**0.748**	**0.604**	**0.562**	**0.582**

Table 3: Ablation experiments on hierarchical difficulty splits (P: Precision, R: Recall, F: F1 Score)

The comparison between "our", "w/o agent", and "ERCA w/o entrance" reveals the distinct roles of the multi-agent mechanism and the entrance model. The "w/o agent" variant replaces the three-agent collaborative verification with a single LVLM, leading to a noticeable decline—especially on hard samples where hard-F drops from 0.582 to 0.554. This confirms that multi-agent voting effectively filters individual model errors and hallucinations in challenging cases. The "ERCA w/o entrance" variant removes the entrance model, routing all samples through the full multi-agent pipeline. Interestingly, its performance slightly exceeds that of "w/o agent" , as the multi-agent collaboration still provides robustness benefits. However, it consistently underperforms the full "our", because the entrance model enables adaptive routing: simple samples benefit from direct single-model inference without the noise introduced by unnecessary multi-agent discussion, while complex samples still leverage the full verification pipeline. This design achieves the optimal balance between accuracy and efficiency, as discussed in Section 4.4.

To further understand the multi-agent dynamics, we analyzed the inter-agent agreement across difficulty levels. On simple samples, agents exhibit high agreement, while on hard samples, agreement drops substantially , indicating that hard cases indeed introduce genuine ambiguity that necessitates the multi-agent voting mechanism.

5 Conclusion

In this paper, we propose ERCA, a multi-modal false information detection framework based on elemental reconstruction and consistency analysis. This framework forms a collaborative multi-agent system by integrating LLMs and LVLMs, which conducts fine-grained consistency analysis at the element level based on the constraints of the original text. We design a dedicated dataset focused on consistency analysis to prove the effectiveness of the aforementioned method. Additionally, ablation experiments have been conducted to study the impact of different modules of the framework on the performance of the method.

In the future, we will further enhance the robustness and generalization ability of the framework by extending the modal and introducing external knowledge, thereby better addressing the unique challenges of multi-modal false information detection.

Limitations

Although ERCA has achieved good performance in multimodal fake news detection, these experiments

are based on Chinese and English datasets. Moreover, the multi-agent architecture, although mitigated by the entry model, still incurs higher computational overhead compared to the single-channel LVLM approach. Secondly, the construction of the dataset relies on the controlled modification of the image, and the element extraction module is limited to five predefined semantic types. This may not fully capture the diversity of real-world situations.

Ethics

In this study, we invited 3 volunteers to annotate and verify some of the data. Each volunteer was paid 10 yuan for each sample. The dataset and model used in this study were sourced from open-access repositories and strictly followed their open-source license agreements. These resources were only used for research purposes and did not violate their open-source terms. This study utilized the API provided by OpenAI and some open-source models for experiments and text polishing.

References

Shuai Bai, Ke qin 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*, abs/2502.13923.

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*, WWW '22, page 2897–2905, New York, NY, USA. Association for Computing Machinery.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Abhimanyu Dubey and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiahui Geng, Jonathan Tonglet, and Iryna Gurevych. 2025. M4fc: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset. *ArXiv*, abs/2510.23508.

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang Zhao. 2025. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):228–236.

Ping He, Hanxue Zhang, Shufu Cao, and Yali Wu. 2025. Cross-modal fake news detection method based on multi-level fusion without evidence. *Algorithms*, 18(7).

Rui Hong and Shuxue Quan. 2026. To see or to please: Uncovering visual sycophancy and split beliefs in vlms. *Arxiv*, abs/2603.18373.

Jiacheng Huang, Tao Han, Ning Yu, and Xiaoyin Yi. 2026. Socratic elenchus-inspired multi-agent debate for mitigating hallucinations in large language models. *Expert Systems with Applications*, 320:132208.

Subin Huang, Qing Zhou, Daoyu Li, Chao Kong, and Sanmin Liu. 2025. Fnd-eka: hierarchical conditional multimodal fake news detection with external knowledge augmentation. *Knowl. Inf. Syst.*, 67(11):10611–10637.

W. Jin, N. Wang, T. Tao, and 1 others. 2024. A veracity dissemination consistency-based few-shot fake news detection framework by synergizing adversarial and contrastive self-supervised learning. *Scientific Reports*, 14:19470.

Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2023. Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, 60(1):103120.

Tanveer Khan, Antonis Michalas, and Adnan Akhunzada. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, 190:103112.

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*, WWW '19, page 2915–2921, New York, NY, USA. Association for Computing Machinery.

Jewon Lee, Ki-Ung Song, Seungmin Yang, Donguk Lim, Jaeyeon Kim, Wooksu Shin, Bo-Kyeong Kim, Yong Jae Lee, and Tae-Ho Kim. 2025. Efficient llama-3.2-vision by trimming cross-attended visual features. *ArXiv*, abs/2504.00557.

X. Li, S. Wang, S. Zeng, and 1 others. 2024a. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(9).

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024b. Large language model agent for fake news detection. *ArXiv*, abs/2405.01593.

Yangfu Li, Hongjian Zhan, Jiawei Chen, Yuning Gong, Qi Liu, and Yue Lu. 2026. Deepscan: A training-free framework for visually grounded reasoning in large vision-language models. *Arxiv*, abs/2603.03857.

Zihao Li, Jiaxin Yang, Xianghan Wang, Jun Lei, Shuohao Li, and Jun Zhang. 2025. Uncertainty-aware disentangled representation learning for multimodal fake news detection. *Inf. Process. Manage.*, 62(5).

Daizong Liu, Wangqin Liu, Xiaowen Cai, Pan Zhou, Runwei Guan, Xiaoye Qu, and Bo Du. 2026. Generating transferable attacks across large vision-language models using adversarial deformation learning. *Pattern Recognition*, 176:113194.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 504–514, New York, NY, USA. Association for Computing Machinery.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongzhen Lv, Wenzhong Yang, Yabo Yin, Fuyuan Wei, Jiaren Peng, and Haokun Geng. 2025. Mdf-fnd: A dynamic fusion model for multimodal fake news detection. *Knowledge-Based Systems*, 317:113417.

Sihan Ma, Qiming Wu, Ruotong Jiang, and Frank Burns. 2025. Contextguard-lvlm: Enhancing news veracity through fine-grained cross-modal contextual consistency verification. *ArXiv*, abs/2508.06623.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, Bangkok, Thailand. Association for Computational Linguistics.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3343–3347, New York, NY, USA. Association for Computing Machinery.

Maged Nasser, Noreen Izza Arshad, Abdulalem Ali, Hitham Alhussian, Faisal Saeed, Aminu Da'u, and Ibtehal Nafea. 2025. A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering*, 26:104752.

Jenny Schmalfuss, Nadine Chang, Vibashan VS, Maying Shen, Andrés Bruhn, and Jose M. Alvarez. 2025. Parc: A quantitative framework uncovering the symmetries within vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25081–25091.

Xiaorong Shen, Maowei Huang, Zheng Hu, Shimin Cai, and Tao Zhou. 2024. Multimodal fake news detection with contrastive learning and optimal transport. *Frontiers in Computer Science*, Volume 6 - 2024.

Peng Shi and Bing Yan. 2021. A survey on intelligent control for multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):161–175.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8:171 – 188.

Kaijia Tian, Guozheng Rao, Xin Wang, Mufan Yu, Jiayin Zhang, and Li Zhang. 2025. Cmfnthinker: A novel cross-source multi-modal fake news detection model. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *ArXiv*, abs/2501.06322.

Ashwin Verma, Soheil Mohajer, and Behrouz Touri. 2025. Multi-agent fact checking. *ArXiv*, abs/2503.02116.

Jinghong Wang, Hongbo Yang, Xizhao Wang, Wei Wang, and Yanan Li. 2026. Magnet: A multimodal knowledge-augmented graph network for early-stage misinformation detection. *Neurocomput.*, 658(C).

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. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yihao Wang, Zhong Qian, and Peifeng Li. 2025. Fmnv: A dataset of media-published news videos for fake news detection. In *Advanced Intelligent Computing Technology and Applications: 21st International Conference, ICIC 2025, Ningbo, China, July 26–29, 2025, Proceedings, Part V*, page 321–332, Berlin, Heidelberg. Springer-Verlag.

Bingbing Xie, Xiaoxiao Ma, Xue Shan, Amin Beheshti, Jian Yang, Hao Fan, and Jia Wu. 2025. Multiknowledge and llm-inspired heterogeneous graph neural network for fake news detection. *IEEE Transactions on Computational Social Systems*, 12(2):682–694.

Danni Xu, Shaojing Fan, Harry Cheng, and Mohan Kankanhalli. 2025a. Rw-post: Auditable evidence-grounded multimodal fact-checking in the wild. *ArXiv*, abs/2512.22933.

Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. 2025b. Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty. *ArXiv*, abs/2510.25120.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

Shuo Yang, Zijian Yu, Zhenzhe Ying, Yuqin Dai, Guoqing Wang, Jun Lan, Jinfeng Xu, Jinze Li, and Edith C. H. Ngai. 2025. Rama: Retrieval-augmented multi-agent framework for misinformation detection in multimodal fact-checking. *ArXiv*, abs/2507.09174.

Hongzhu Yu, Hongchen Wu, Xiaochang Fang, Meng Li, and Huaxiang Zhang. 2025. Sr-cibn: Semantic relationship-based consistency and inconsistency balancing network for multimodal fake news detection. *Neurocomput.*, 635(C).

Zhi Zeng, Yifei Yang, Jiaying Wu, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, and Minnan Luo. 2026. From manipulation to mistrust: Explaining diverse micro-video misinformation for robust debunking in the wild. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 7621–7632, New York, NY, USA. Association for Computing Machinery.

Guanghao Zheng, Bowen Shi, Mingxing Xu, Ruoyu Sun, Peisen Zhao, Zhibo Zhang, Wenrui Dai, Junni Zou, Hongkai Xiong, Xiaopeng Zhang, and Qi Tian. 2025. Granvit: A fine-grained vision model with autoregressive perception for mllms. *ArXiv*, abs/2510.21501.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).