HybridNet: Efficient Multimodal Fake News Detection *

Shreyas Kumar Tah, Anshul Singh, Prajeet Katari, Aditya Agarwala, Shwetabh Biswas, Lucky Gupta, Siddhartha Banerjee*, Faheema AGJ*, Ashika*, Soma Biswas
Indian Institute of Science, Bangalore, India *CAIR, DRDO, India
siddhart.cair@gov.in, faheema.cair@gov.in, ashika.cair@gov.in
shreyaskumartah@gmail.com, katariprajeet26@gmail.com
anshulsinghchambial@gmail.com, shwetabhbiswas2305@gmail.com
luckygupta@iisc.ac.in, aditaa1@iisc.ac.in, somabiswas@iisc.ac.in

Abstract

Detecting multimodal fake news, where authentic images are paired with misleading text, remains a significant challenge. Vision Language Models give impressive performance for this task, but the results are not easily interpretable. Fine-tuning large multimodal LLMs can provide explanations, but relies heavily on large, fully annotated datasets with reasoning to achieve high accuracy, ultimately constraining its real-world scalability. In this work, we propose HybridNet, treating frozen open-source MLLMs as reasoning extractors rather than classifiers through three-stage consistency-checking. We distill these signals into a lightweight Reasoning-Aware Classifier that weighs MLLM observations against VLM features, enabling it to move beyond binary predictions and produce interpretable explanations alongside final outputs. In addition, HybridNet employs an active learning strategy to combine MLLM interpretability with supervised robustness, achieving competitive accuracy using less than 50% *labeled data and offering a scalable and interpretable solution for multimodal misinformation detection.*

1. Introduction

Multimodal Fake News Detection (MFND) has become a critical task in the moderation of social media, as misinformation is often spread by pairing authentic images with misleading or unrelated text [4]. Recent research in MFND has followed two distinct paths. On one hand, Vision-Language Models (VLMs) like CLIP [1, 9, 11] are highly efficient and obtain high accuracy by aligning visual and textual features to identify fake news. However, this approach fails to provide human-interpretable explanations of their decision-making process, which fact-checkers and users need to understand why a post is flagged. On the other hand, Multimodal Large Language models (MLLMs) can generate detailed reasoning through Chain-of-Thought prompting [6, 10]. Yet, these models are computationally expensive to fine-tune. Fine-tuning also requires large datasets with high-quality explanations, collecting which is a labor-intensive task that requires expert knowledge and is often very expensive to scale.

Here, we propose **HybridNet**, a novel framework designed to provide both high accuracy and interpretability without the need for large-scale reasoning datasets or expensive MLLM fine-tuning. HybridNet utilizes a frozen open-source MLLM as a reasoning guide. Our approach centers on the idea that we can distill the reasoning capabilities of an MLLM into a lightweight classifier that works alongside the features of VLM-based models. Unlike prior approaches that employ MLLMs either as end-to-end classifiers or explanation generators, HybridNet treats a frozen MLLM strictly as a structured reasoning extractor. The model does not rely on MLLM predictions for final decision-making. Instead, intermediate consistency signals are algorithmically distilled into a trainable classifier that learns how to selectively leverage reasoning cues in conjunction with vision-language features. This shifts the role of MLLMs from decision-makers to supervisory signal generators, enabling scalable and label-efficient learning.

Towards this goal, we introduce a *three-stage consistency-checking process for the MLLM*, which analyzes the query and external evidences. We train a *reasoning-aware classifier* that fuses these reasonings with the latent features from a VLM. To address the difficulty of obtaining sufficient labeled data, we employ an *active learning* strategy to effectively choose the most informative training samples. We observe that a hybrid extension of the widely used entropy based uncertainty measure can effectively identify the most informative samples even

*The work is partially funded by DRDO, Ministry of Defence, Government of India, No. DFTM/02/3125/M/03/AIR-03. We are grateful to Director CAIR, Director DIARCOE, and DFTM for their support.

for the multimodal framework. This allows HybridNet to reach comparable performance as the state-of-the-art using less than half the training data of the fully supervised frameworks. Our main contributions are summarized as follows:

- We introduce a three-stage structured consistency decomposition that breaks multimodal verification into image–text, image–image, and text–text alignment signals, producing interpretable and modular reasoning representations.
- We propose a Reasoning-Aware Classifier (RAC) that distills structured MLLM reasoning into a lightweight Transformer-based fusion architecture, enabling supervised integration of reasoning with VLM embeddings.
- We design a hybrid uncertainty-driven active learning strategy leveraging joint uncertainty of BaseNet and RAC, significantly reducing annotation cost while preserving performance.
- We demonstrate that structured reasoning distillation achieves near state-of-the-art accuracy using only 30–50% labeled data.

2. Related Literature

Here, we briefly review the related literature on multimodal fake news detection (MFND).

Training-based MFND Frameworks: Early approaches for MFND treated the task as a binary classification problem (real, fake), utilizing separate encoders for visual and textual modalities. Models such as SpotFake [13] and SAFE [17] employed pre-trained CNNs (like VGG-19) for images and Transformers like BERT for text, fusing these features to detect inconsistencies between the image and text data.

With the advent of large-scale vision-language models (VLM), recent methods have shifted towards CLIP [11]-based architectures to better capture cross-modal alignment. For instance, CCN [1] introduces a cycle-consistency check to verify if the image and text retrieve similar external evidence. Building on this, RedDot [8] and FraudNet [9] incorporate advanced attention mechanisms and evidence fusion strategies to improve detection accuracy on benchmarks like NewsClippings [4]. *While these methods achieve high empirical performance, they operate as "black boxes," offering little insight into why a specific news item is flagged as fake.*

To address this interpretability gap, few recent approaches utilized Multimodal Large Language Models (MLLMs) for this task. SNIFFER [10] fine-tunes an InstructBLIP model using a two-stage process to align visual concepts with news entities, subsequently training on GPT-4 generated explanations. However, relying solely on MLLMs for classification is *computationally expensive and suffers from high latency*. Furthermore, fine-tuning these large models *requires curated reasoning datasets*, which are difficult to obtain compared to standard binary labels.

API-based MFND Frameworks: Several unimodal approaches that works in zero-shot setup augment CoT with persuasion analysis [6] or implements contamination-resistant evaluation [15] for fake news detection to mitigate inflated performance metrics due to data leaked during VLM-based training. More recently, DEFAME [2], an evidence-based multimodal fact-checking method, introduced a zero-shot framework that integrates multimodal LLMs for dynamic evidence collection to address static knowledge limitations. Additionally, certain hybrid frameworks [5] requires training a dedicated module to model event-level inconsistencies or propagates LLM-generated pseudo-labels via graph networks [3].

Majority of these frameworks rely heavily on costly commercial APIs, e.g., GPT-4, etc, and complex multi-step retrieval causing runtime latency. In contrast, *our approach utilizes an open-source MLLM pipeline specifically for generating intermediate reasoning signals, which are then distilled into a lightweight classifier, avoiding the high recurring costs of API-based verification.*

3. Problem Statement

A multimodal news article is represented as $\{I_q, T_q\}$, where I_q and T_q denote the query image and query text respectively. The task is to predict its veracity label $y_q \in \{\text{true}, \text{fake}\}$. Optionally, external evidences may be available, where visual evidences obtained using query text search are denoted by I_k^e, and textual evidences obtained using query image search are denoted by T_k^e. Here, $k = 1, \ldots, N_e$, where N_e is the number of evidences. Given $\{I_q, T_q, y_q\}$ and optionally $\{I_k^e, T_k^e\}$, the goal is to learn a model f_θ that can correctly predict the veracity of an unseen multimodal news along with human-interpretable explanations.

Towards this goal, we develop a novel HybridNet framework, which addresses the following challenges: (i) limited interpretability of vision-language models, which hinders real-world adoption; (ii) the high computational and data requirements for training MLLM-based approaches that can address interpretability; and (iii) the scarcity of large, high-quality annotated datasets for the MFND task.

4. Proposed HybridNet Framework

Our proposed framework, HybridNet (Figure 1), introduces a data-efficient training paradigm that synergizes a state-of-the-art Vision-Language Model (VLM) with the reasoning capabilities of a Multimodal Large Language Model (MLLM). It consists of three modules, namely: 1) *MLLM-*

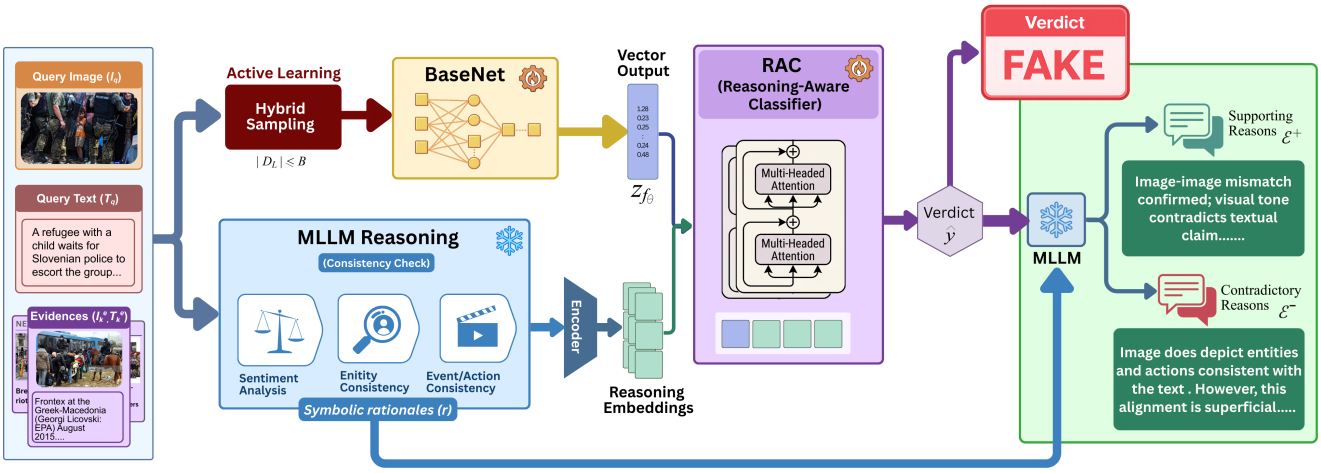


Figure 1. Overview of **HybridNet**. The framework combines BaseNet multimodal feature embedding $\mathbf{z}_{f_\theta}$ with MLLM-generated reasoning embeddings $(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}})$ using the lightweight Reasoning-Aware Classifier (RAC). RAC produces the final prediction $\hat{y}$, and the MLLM generates supporting and contradicting explanations $(\mathcal{E}^+, \mathcal{E}^-)$ conditioned on $\hat{y}$.

based Reasoning Module which performs 3-stage consistency checking on a given news sample, in a training - free manner; 2) *Reasoning Aware Classifier*, which is a lightweight module tuned to perform effectively on difficult samples and 3) *Active learning module* to select the most informative samples for training, thereby significantly reducing high annotation costs. This multi-stage approach significantly reduces dependency on large-scale labeled datasets while maintaining high detection accuracy and providing interpretability.

4.1. MLLM-based Reasoning Module

Directly querying an MLLM to classify a multimodal news as fake or real often leads to hallucinations due to the complex interplay of the modalities. Rather than relying on a single long-form judgment, we constrain the reasoning process to localized and verifiable consistency comparisons across modality pairs. Decomposing the task into **Image-Text**, **Image-Image**, and **Text-Text** alignment reduces open-ended generation and limits hallucination by anchoring reasoning to observable evidence. These structured outputs are not directly trusted for prediction; instead, they serve as intermediate supervisory signals for downstream learning.

We begin by formulating a training-free MLLM pipeline that, given a query pair (I_q, T_q) and its corresponding external evidences $\{(I_k^e, T_k^e)\}_{k=1}^{N_e}$, performs three-stage consistency checking while generating reasoning embeddings $(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}})$. The stages proceed sequentially: first between query image I_q and text T_q, then I_q and visual evidences $\{I_k^e\}_{k=1}^{N_e}$, finally T_q and textual evidences $\{T_k^e\}_{k=1}^{N_e}$. At each stage, our prompt-based pipeline extracts structured reasoning on sentiment, entity, and event alignment across image-text, image-image, and text-text pairs, capturing consistency/ambiguity signals for downstream classification. More details of the prompts can be found in Appendix 8.

Image-Text ($\mathbf{r}_{\text{img-txt}}$): Here, the MLLM is tasked to analyze the cross-modal consistency of the query image and its caption. For the refugee/police news in Figure 2 (GT: FAKE), MLLM identifies: Sentiment *Mismatch* ('neutral text vs image anxiety/vulnerability'), Entities *Aligned* ('refugee, child, police present'), Event/Action *Aligned* ('police escort scenario'). Despite entity/event support, sentiment discrepancy flags potential misrepresentation, providing interpretable evidence of cross-modal issues.

Image-Image ($\mathbf{r}_{\text{img-img}}$): Here, the query image is compared with the top-ranked retrieved evidence image. For the example in Figure 2, the MLLM finds: Sentiment *Aligned* ('negative: tense police/refugee scenes'), Entities *Aligned* ('police, refugees/migrants in both'), Event/Action *Aligned* ('border control/containment'). 'All analyses indicate alignment' validating visual similarity as both images depict similar situations involving police and refugees/migrants, suggesting a consistent narrative

Text-Text ($\mathbf{r}_{\text{txt-txt}}$): Here, the MLLM evaluates the veracity of the query caption by comparing with the retrieved evidence captions. For the sample in Figure 2, MLLM categorizes all evidence captions as contradictory, because each of them focuses on different locations (Greek-Macedonia border) and time periods, demonstrating a lack of factual alignment with the original query caption. The framework gives more weightage to the evidence that is obtained from any verified sources.

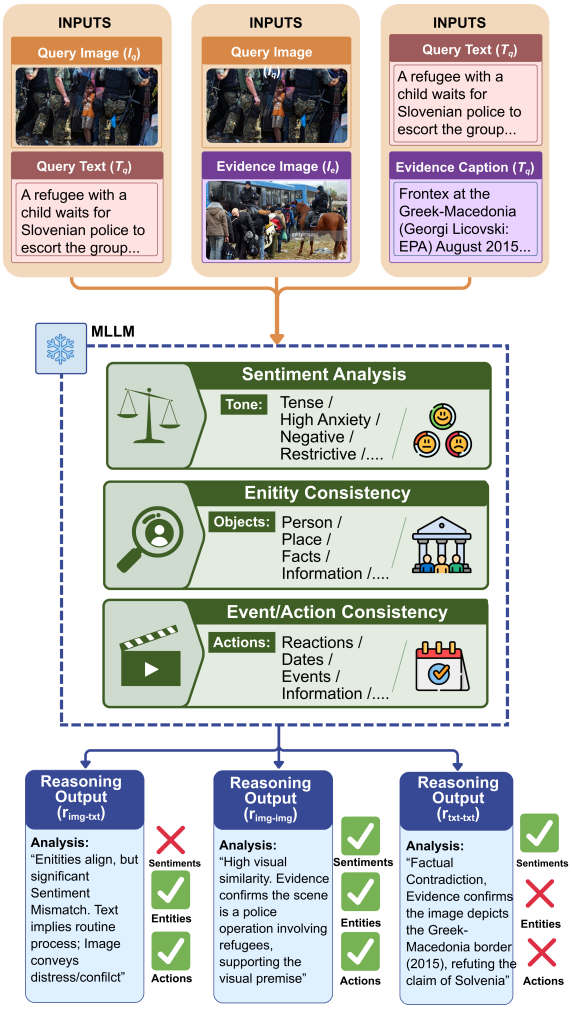


Figure 2. Illustration of the proposed three-stage consistency-checking pipeline. (1) Image–Text consistency evaluates semantic alignment between the image and its caption across sentiment, entity presence, and event/action depiction to assess whether the textual claim is supported by the visual content. (2) Image–Image consistency compares the query image with retrieved visual evidence to verify whether they depict the same real-world situation. (3) Text–Text consistency assesses factual agreement between the query caption and retrieved textual evidence. By jointly reasoning over these complementary consistency signals, the pipeline strengthens confidence in the final decision. For instance, if the query image fails to align with the retrieved visual evidence obtained using the caption, this mismatch indicates a potential image–text inconsistency, raising a strong cue for fake news.

The proposed training-free MLLM pipeline produces high-quality and human-interpretable reasoning outputs, but struggles to integrate them properly to give a correct real/fake prediction, making the classification unreliable. Therefore, we utilize the MLLM explanations $M(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}})$ as structured supervisory signals and integrate them with a VLM-based BaseNet architecture [9, 11], FraudNet [9].

4.2. Reasoning-Aware Classifier

This lightweight RAC module combines the benefits of both the models - high performance of the VLM from supervised fine-tuning and fine-grained explanation of the frozen MLLM and improves robustness on difficult samples while maintaining interpretability. First, the BaseNet's feature embedding (final multimodal embedding used for classification), $\mathbf{z}_{f_\theta}$, is combined with the embeddings of the three reasonings obtained by the MLLM. Specifically, the reasonings $\mathbf{r}_{\text{img-txt}}$, $\mathbf{r}_{\text{img-img}}$, $\mathbf{r}_{\text{txt-txt}}$ are converted into a fixed-dimensional vector representation via a reasoning encoder ψ as: $\psi_{img-txt}$, $\psi_{img-img}$ and $\psi_{txt-txt}$. Next, these representations and $\mathbf{z}_{f_\theta}$ are projected into a common latent space.

$$\begin{aligned} \mathbf{h}_j &= \text{Proj}_j(\mathbf{k}_j), \\ \mathbf{k}_j &\in \left\{ \mathbf{z}_{f_\theta}, \psi_{img-txt}, \psi_{img-img}, \psi_{txt-txt} \right\} \end{aligned} \tag{1}$$

The projected embeddings are combined to form a sequence with a learnable classification token $\mathbf{e}_{\text{cls}}$: $\mathbf{S} = \left[\mathbf{e}_{\text{cls}}, \mathbf{h}_{f_\theta}, \mathbf{h}_{\text{img-txt}}, \mathbf{h}_{\text{img-img}}, \mathbf{h}_{\text{txt-txt}}\right]$. This is finally processed by a Transformer Encoder, and the `[CLS]` output is passed to an MLP head to output the final classification. Importantly, RAC does not treat reasoning and BaseNet embeddings as static parallel inputs. Instead, self-attention over the joint token sequence enables reasoning cues to dynamically modulate the shared representation via the classification token. This allows adaptive weighting of consistency signals, where BaseNet features dominate simpler samples and reasoning tokens exert greater influence in ambiguous cases, leading to robust performance gains.

To analyze the contribution of each reasoning cue, we conduct an ablation for RAC with individual signals. With 30% labeled data, full integration achieves 90.33% accuracy versus 87.69% (image-text), 88.24% (image-image), and 88.50% (text-text) alone, indicating that gains stem from complementary multi-cue interactions rather than simple concatenation.

In this way, structured reasoning acts as controllable soft evidence rather than fixed auxiliary features, improving robustness while preserving interpretability.

4.3. Active Learning Module

Since verifying a news aticle as true or fake requires significant amount of manual labor and domain expertise (eg. a doctor can reliably infer the veracity of a medical news), here we explore whether active learning (AL) can be used to identify the informative samples to be annotated, thereby reducing annotation costs. Given the unlabeled training data, we start with a small, randomly sampled annotated seed set D_L. We propose a hybrid sampling AL strategy to iteratively augment D_L by selecting the most informative samples from the remaining unlabeled pool D_U, till we reach

the predefined annotation budget B. Specifically, we select samples to prepare the RAC for the difficult cases where the BaseNet is less confident. We employ an intersection-based strategy and annotate only samples that are highly uncertain for *both* models (BaseNet and RAC) to target cases where *both* models struggle with complementary weaknesses. The AL strategy is formally described in the joint training strategy of HybridNet, which is detailed next.

5. Training and Inference

Here, we describe the training and inference strategy of the proposed HybridNet framework.

Training Stage: In each iteration, we first train the BaseNet f_θ on the labeled set D_L. For hyperparameter tuning, the full validation set D_V is used. Simultaneously, to prepare supervision for RAC tuning, we identify low-confidence samples from D_V where the BaseNet underperforms:

$$D_V^{\text{low}} = \{x_j \in D_V \mid H(f_\theta(x_j)) > \tau\}. \tag{2}$$

This ensures that RAC is trained to improve performance, specifically on hard samples where the BaseNet is not confident.

Once the BaseNet is trained, the RAC g_ϕ is trained on the same labeled set D_L, but tuned primarily on the BaseNet uncertain validation set D_V^{low}, leveraging both the BaseNet feature representation $\mathbf{z}_{f_\theta}$ and MLLM-generated reasoning embeddings $(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}})$ to improve performance on the hard samples. Since RAC is trained with ground-truth supervision, it learns to suppress noisy or inconsistent reasoning signals during optimization. Through attention-based fusion, unreliable reasoning embeddings can be down-weighted in favor of stronger BaseNet features. This supervised filtering mechanism mitigates potential MLLM hallucinations and contributes to the observed performance gains over standalone reasoning.

For each unlabeled sample $x_j \in D_U$, BaseNet and RAC independently compute predicted probabilities $p_j = f_\theta(x_j)$ and $q_j = g_\phi(x_j)$, and uncertainty is quantified using Shannon entropy:

$$H(p_j) = -p_j \log_2(p_j) - (1-p_j)\log_2(1-p_j),$$
$$H(q_j) = -q_j \log_2(q_j) - (1-q_j)\log_2(1-q_j).$$

Both models select a preliminary set of highly uncertain samples, and the final batch D_{select} is chosen as their intersection.

Only D_{select} is annotated and added to D_L, after which both BaseNet and RAC are retrained on the expanded labeled set. This procedure repeats until the annotation budget B is reached. We also experiment with selecting the union of the preliminary sets, and find that the intersection strategy yields comparable performance with fewer annotated samples, making it more sample efficient.

Table 1. Zero-shot performance of MLLMs on NewsClippings using our three-stage consistency checking (Img-Txt: Image-Text alignment, Img-Img: Image-Image evidence matching, Txt-Txt: Text-Text fact verification). Each intermediate consistency score is obtained by deriving a predicted label from the respective reasoning stage and comparing it against the original ground truth. Gemma demonstrates superior reasoning across all stages.

MLLM Backbone	**Img-Txt (%)**	**Img-Img (%)**	**Txt-Txt (%)**	**Overall Acc. (%)**
InternVL-3	64.8	71.7	58.5	58.7
Qwen2.5-VL	63.0	72.5	53.8	52.3
Gemma (Ours)	**68.7**	**73.8**	**59.2**	58.2

Interpretable Inference Stage: During inference, the RAC produces the final output {real, fake}, by conditioning on both $\mathbf{z}_{f_\theta}$ and the MLLM-derived reasoning representations, thereby retaining accuracy while remaining explicitly grounded in the three-stage consistency signals, as shown in Figure 1. To provide human-interpretable justifications for the RAC prediction, we feed the predicted label together with the original consistency reasonings into the MLLM to obtain explanations *supporting* and *contradicting* the RAC decision. Concretely, given the RAC prediction $\hat{y}$, we compute:

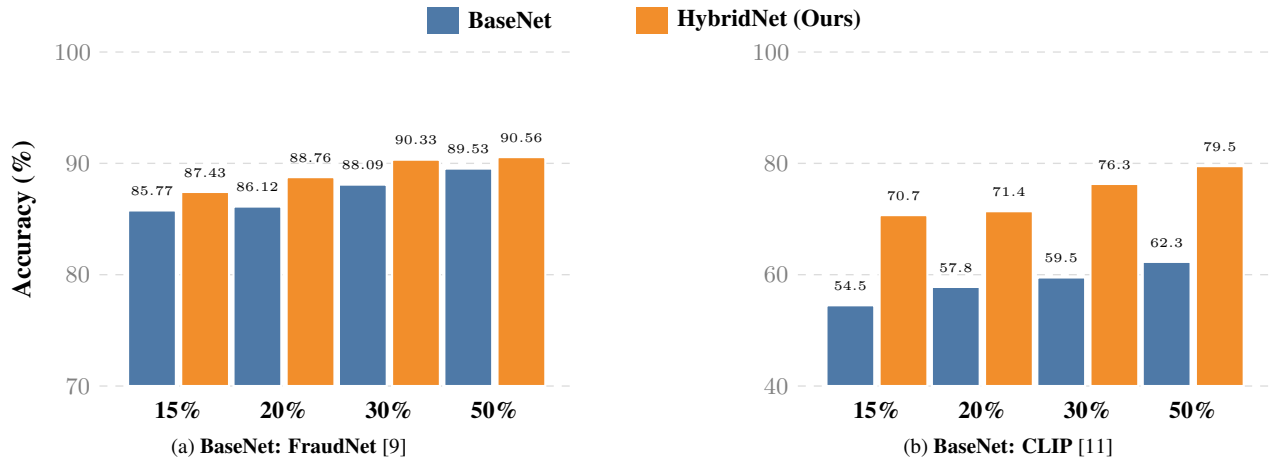
$$(\mathcal{E}^+, \mathcal{E}^-) = MLLM\big(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}}, \hat{y}\big)$$

where $\mathcal{E}^+$ and $\mathcal{E}^-$ denote the sets of reasons in favor of and against the predicted label, respectively. This design makes HybridNet a general plug-in mechanism for augmenting strong MFND backbones that lack human-readable explanations, while maintaining its performance.

6. Experimental Evaluation

Here, we conduct extensive experiments to rigorously evaluate the proposed framework.

Datasets Used: A popular benchmark for multimodal fake news detection, the **NewsClippings** [4] dataset, is used for all experiments. It contains real-world image–caption pairs, where fake samples are synthetically created by pairing real captions with images from different articles to form out-of-context mismatches. We follow the standard splits with 71k training, 7k validation, and 7k test samples. For cross-dataset evaluation, we utilize the **IFND** dataset [12], which contains originally curated fake from various fact-checked sources, and **Tampered News** [7], which contains manually manipulated media distinct from those in synthetically created Newsclippings.



(a) **BaseNet: FraudNet** [9] (b) **BaseNet: CLIP** [11]

Figure 3. **Data Efficiency:** Performance of HybridNet (Orange) versus BaseNet (Blue) for varying percentages of training data using the same samples selected using the hybrid AL strategy. **(a)** With the FraudNet backbone, our method surpasses the baseline with consistently higher accuracy, achieving near-saturation at just 30% data. **(b)** With the standard CLIP backbone, HybridNet demonstrates substantial gains ($> 15\%$), highlighting the robustness of our reasoning-distilled approach in low-data regimes.

Implementation Details: We evaluate HybridNet on two BaseNets, FraudNet [9] and original CLIP ViT-L/14 [11] with MLP head. Active learning starts with a randomly sampled 10% seed set, fixed across all experiments. Performance is evaluated at annotation budgets of **15%**, **20%**, **30%**, and **50%**. In our hybrid strategy, we use $\alpha = 2.5\%$ of the unlabeled pool for batch size per iteration. Training proceeds for up to 30 epochs, with early stopping (patience=3) when both models fail to improve for three consecutive epochs. RAC is a lightweight transformer (hidden size 512, 2 attention heads, dropout 0.3), trained with Adam (1×10^{-4}) and early stopping (patience 5) on validation accuracy. The reasoning embeddings $\psi()$ uses pre-trained Qwen [16]. For cross-data generalization, we use HybridNet with FraudNet BaseNet trained on 50% samples. Fine-tuning follows the same hybrid sampling with 5 epochs.

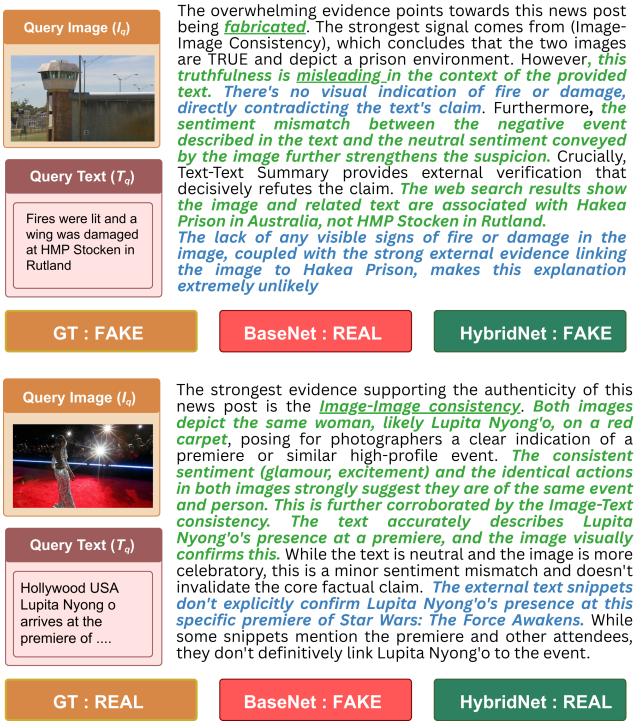
6.1. Choice of MLLM

To evaluate how MLLM selection impacts MFND performance, we tested three open-source MLLMs on the NewsClippings [4] dataset. We instruct the model to independently evaluate semantic dimensions, specifically *Sentiment Alignment*, *Entity Consistency*, and *Event/Action Depiction* before aggregating these insights into a final verdict. This structured breakdown forces the model to ground its decision in observable evidence rather than abstract biases, providing high-quality embeddings for the RAC.

Table 1 reveals MLLMs' challenges in final judgment: despite strong individual analyses (Gemma3: 68.7% Img-Txt, 73.8% Img-Img, 59.2% Txt-Txt), overall accuracy is comparatively lower (58.2%). This shows that MLLMs struggle to effectively combine the explanations to form a robust judgement. Gemma3 [14], with a higher parameter count than medium-sized InternVL-3 [18] and Qwen2.5-VL [16], consistently outperformed others due to superior *world knowledge* and *regional knowledge*, essential for identifying context-specific entities in diverse news narratives, and shows superior branch-wise reasoning. We select Gemma3 (12B) as HybridNet's reasoning backbone for its robust intermediate analysis. More details on prompts struc-

Table 2. Comparison of the proposed HybridNet with state-of-the-art approaches. BaseNet refers to either FraudNet [9] or CLIP [11] backbone. We observe that HybridNet trained on 30-50% data (with FraudNet as base network) performs at par with the SOTA frameworks with 100% training data on NewsClippings dataset by integrating MLLM reasoning with active learning.

Method	Data (%)	Acc. (%)
RANDOM SAMPLING		
CLIP BaseNet	30	57.7
CLIP BaseNet	50	62.1
FraudNet BaseNet	30	86.9
FraudNet BaseNet	50	87.8
HYBRIDNET (PROPOSED)		
CLIP HybridNet	**30**	**76.28**
CLIP HybridNet	**50**	**79.45**
FraudNet HybridNet	**30**	**90.33**
FraudNet HybridNet	**50**	**90.56**
FULLY SUPERVISED		
CLIP [11]	100	68.7
CCN [1]	100	84.7
Sniffer [10]	100	88.4
RedDot [8]	100	90.3
FraudNet [9]	100	91.1



(a) Challenging examples where BaseNet fails but HybridNet correctly classifies along with explanations.

(b) Examples where predictions of BaseNet & HybridNet agree. (Top): Correct and (Bottom): Incorrect Prediction.

Figure 4. **Qualitative examples of HybridNet on Newsclippings** dataset along with the explanations. Green color indicates supporting reasonings ($\mathcal{E}^+$) while those in blue color refers to reasonings against ($\mathcal{E}^-$) the HybridNet prediction.

ture and reasoning outputs are provided in Appendix 8 and Appendix 9.

6.2. Experimental Results

While open-source MLLMs generate high-quality, context-rich reasoning through structured prompting, they often hallucinate when producing final veracity judgments under long, diverse contexts. This unreliability motivates HybridNet as a hybrid solution: we distill the MLLM's three-stage consistency rationales $(\mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{txt-txt}})$ into a lightweight Reasoning-Aware Classifier (RAC) that couples them with BaseNet features $\mathbf{z}_{f_\theta}$, achieving both robust classification and grounded explanations.

Table 2 compares the proposed HybridNet with the state-of-the-art approaches for the MFND task. The results for the other approaches using 100% training data are taken directly from [9]. We make the following observations:
(1) Effectiveness of the hybrid AL strategy: We observe that the proposed, hybrid AL strategy significantly outperforms random sampling, achieving the highest accuracy across different budgets, confirming its suitability for sample-efficient training using our proposed framework;
(2) Data efficiency : Using only 30% training data, HybridNet (with FraudNet as BaseNet) achieves an accuracy of 90.33%, which is comparable to the state-of-the-art using 100% training data, justifying its effectivess as a *data efficient and interpretable MFND framework*. HybridNet is designed to improve robustness under limited supervision rather than replace fully saturated classifiers.The gains are most pronounced in low-data regimes, where reasoning-guided learning compensates for insufficient annotations.
(3) Effectiveness of RAC : Figure 3 shows HybridNet's performance compared to BaseNet on the same set of samples. With FraudNet as BaseNet, at $\sim 50\%$ training data, HybridNet reaches **90.56%** (vs. BaseNet 89.53%). With CLIP as BaseNet, gains are larger: from 62.3% to **79.45%**. The proposed sampling strategy enables HybridNet to accelerate convergence to near-supervised performance with fewer data.

When trained with 100% labeled data, FraudNet operates near performance saturation. In this setting, HybridNet remains competitive (91.92% vs. 91.1%), indicating that reasoning integration does not degrade strong baselines, while its primary advantage emerges in reduced-label and low-confidence scenarios.

Importantly, HybridNet avoids expensive MLLM fine-tuning. A frozen 12B-parameter MLLM is used solely for structured reasoning extraction, while the RAC module contains approximately 5.4M parameters. On a single A6000 GPU, inference requires approximately 5 min-

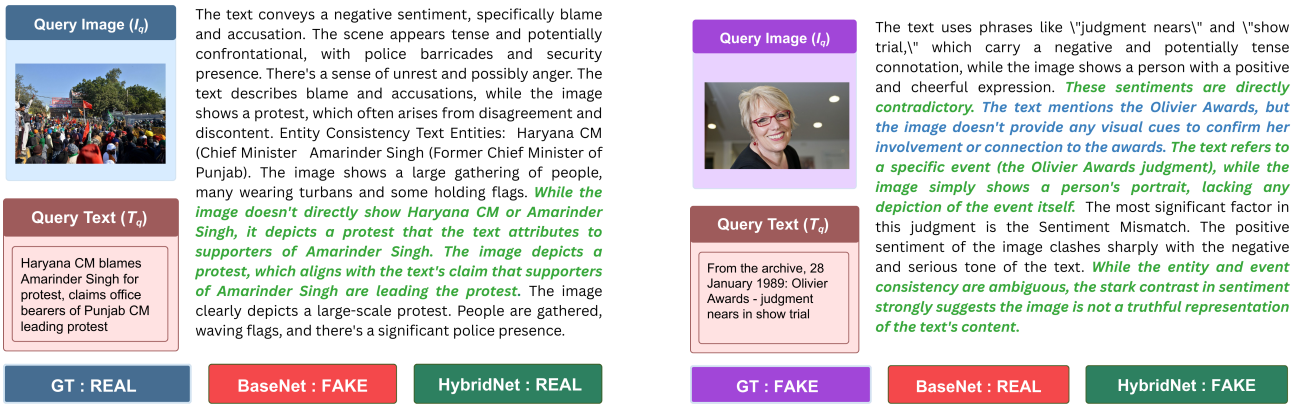


Figure 5. **Qualitative examples of HybridNet on IFND (left) and Tampered (right)** dataset along with the explanations. Green color indicates supporting reasonings ($\mathcal{E}^+$) while those in blue color refers to reasonings against ($\mathcal{E}^-$) the HybridNet prediction.

utes per 100 samples, making the framework practical for moderate-scale deployments.

Qualitative Results Figure 4 (a) shows two challenging examples where BaseNet [9] misclassifies, but HybridNet corrects it using MLLM reasoning cues; (b) illustrates two examples where both the models agree. The top example is an easy one, where both models are correct, and the bottom one is a difficult example, where both models fail.

6.3. Cross-Data Generalization

We evaluate HybridNet trained on NewsClippings (71k train/7k valid/7k test) [4] at 50% of data using the proposed method for cross-dataset generalization on IFND News (2.5k train/1.3k valid/16k test) [12] and Tampered News datasets (2.1k train/1.3k valid/16k test)[7]. HybridNet achieves 90.4/88.2% overall accuracy on test sets (IFND/Tampered columns) when finetuned on less than 50% of the data (Figure 6), substantially improving over zero-shot baselines (55-64%) and reaching near full fine-tuning. This justifies the effectiveness of HybridNet and the AL strategy to quickly adapt to other datasets as shown in Figure 5. Note that these datasets do not contain evidences.

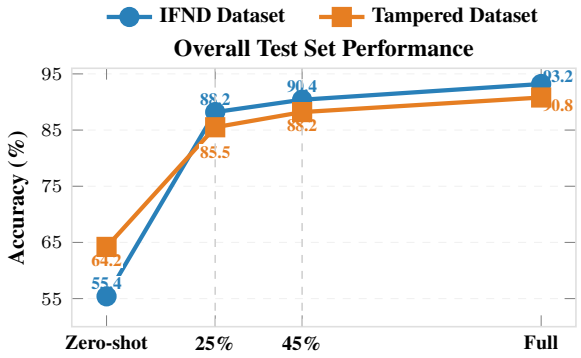


Figure 6. **Cross-dataset generalization on IFND and Tampered datasets.** Overall test performance improves rapidly with HybridNet using 25–45% training data.

7. Conclusion

Here, we propose a novel framework, HybridNet, for the challenging real-world task of label-efficient multimodal fake news detection. It leverages the powerful reasoning capabilities of open-source MLLMs using a three-stage consistency checking mechanism. This training-free pipeline removes the requirement of collecting huge dataset of news with detailed explanations and computationally intensive fine-tuning of these models. HybridNet seamlessly integrates the state-of-the-art CLIP-based MFND models with the MLLM's structured Chain-of-Thought analysis using a lightweight Reasoning-Aware Classifier. An additional hybrid active learning strategy helps to significantly reduce the annotation cost. Extensive experiments on show that HybridNet not only compares favorably to full-supervision training using less than 50% labeled data, but also generalizes effectively to other datasets. Thus, HybridNet delivers a scalable, interpretable solution for real-world multimodal misinformation detection.

Limitations and Ethical Considerations

Our evaluation is limited to English datasets (NewsClippings [4], IFND [12]). Extending HybridNet to multilingual settings is a natural next step with modern MLLMs. The current MFND formulation assumes authentic images; integrating DeepFake detection remains future work. As misinformation detection has societal impact, HybridNet is intended as a decision-support tool rather than an autonomous system. While we use public datasets, MLLMs may exhibit bias or hallucination, requiring human oversight.

References

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14940–14949, 2022. 1, 2, 6

[2] Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. Defame: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*, 2024. 2

[3] Shuguo Hu, Jun Hu, and Huaiwen Zhang. Synergizing llms with global label propagation for multimodal fake news detection. *arXiv preprint arXiv:2506.00488*, 2025. 2

[4] Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*, 2021. 1, 2, 5, 6, 8

[5] Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 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, 2024. 2

[6] Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, 2025. 1, 2

[7] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 16–25, 2020. 5, 8

[8] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Red-dot: Multimodal fact-checking via relevant evidence detection. *IEEE Transactions on Computational Social Systems*, 2025. 2, 6

[9] Devendra Patel, Vikas Verma, Shreyas Kumar Tah, Shwetabh Biswas, and Soma Biswas. Fraud-net: Fraud news detection using sample uncertainty & domain aware generalized network. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3363–3371. IEEE, 2025. 1, 2, 4, 6, 7, 8

[10] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13062, 2024. 1, 2, 6

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 4, 6

[12] Dilip Kumar Sharma and Sonal Garg. Ifnd: a benchmark dataset for fake news detection. *Complex & intelligent systems*, 9(3):2843–2863, 2023. 5, 8

[13] Srishti Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47. IEEE, 2019. 2

[14] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 6

[15] Cheng Xu and Nan Yan. Triplefact: Defending data contamination in the evaluation of llm-driven fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8808–8823, 2025. 2

[16] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6

[17] Xinyi Zhou, Jingwei Wu, and Reza Zafarani. Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 354–367. Springer, 2020. 2

[18] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 6