HybridNet: LLM-Guided Active Learning for Multimodal Fake News Detection

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

Abstract

*Detecting multimodal fake news (MFND), where authentic images are paired with misleading text, remains a significant challenge. Existing methods achieve high accuracy but rely heavily on large, fully annotated datasets, limiting real-world scalability. We propose **HybridNet**, a data-efficient framework that leverages hybrid active learning to select the most informative samples, drastically reducing labeling cost. We also propose a lightweight **Reasoning-Aware Classifier (RAC)** for challenging cases, which combines Vision–Language Model (VLM) features with reasoning from a Multimodal Large Language Model (MLLM) to further improve detection performance and provide human-interpretable explanations. Our hybrid approach combines uncertainty-based bootstrapping with LLM-guided disagreement to prioritize contextually difficult samples. We further evaluate HybridNet for cross-dataset generalizability, demonstrating that it maintains strong performance across diverse multimodal news corpora. Experiments on benchmark datasets show that HybridNet achieves competitive accuracy with less than half the labeled data, offering a scalable and interpretable solution for multimodal misinformation detection.*

1. Introduction

The field of Multimodal Fake News Detection (MFND) has become a crucial area of study that addresses one of the most prevalent types of digital misinformation, which is the intentionally created pairing of misleading text with authentic images. Research has shown that people are more likely to accept both true and false statements when they are accompanied by images, as photographs not only boost perceived credibility but also drive higher engagement on social media. This makes the task of MFND especially challenging, as it requires not only detecting discrepancies between textual and visual modalities but also capturing the subtle connections that distinguish carefully crafted misinformation from genuine news. In recent years, two major research directions have emerged to address the challenging problem of multimodal fake news detection.

Vision–Language Model (VLM) Approaches : The first direction leverages pre-trained vision–language models, particularly **CLIP** [13] based architectures, to achieve multimodal alignment in shared latent spaces. These methods encode images and accompanying texts into common embedding representations and employ cosine similarity metrics with contrastive learning objectives to assess veracity. Methods like **CCN** [1], **Reddot** [10], and **FraudNet** [11] have extended this paradigm by introducing advanced attention mechanisms, cross-modal consistency losses, and feature fusion strategies to enhance alignment and detection performance. Despite their computational efficiency and strong empirical results, VLM-based systems remain limited by their "black box" nature, offering little interpretability into how predictions are made. This lack of explainability makes it difficult to implement in the real world, particularly when moderators and human fact-checkers need to know why particular information is flagged.

Multimodal Large Language Model (MLLM)-Based Approaches: The second major research trajectory in MFND focuses on MLLM-based methods, which aim to deliver both robust classification and human-interpretable explanations, grounded in external evidence retrieval and reasoning. For instance, **SNIFFER** [12] is a multimodal LLM meticulously fine-tuned via a two-stage instruction tuning process, which first aligns generic visual concepts with news-domain entities, then trains on **GPT-4** [3] generated out-of-context examples to surpass its base MLLM by 40%.

While MLLM-based methods effectively bridge the ex-

plainability gap left by VLM approaches, they come with practical challenges. Both training and inference demand substantial computational resources and time, making them costly and resource-intensive. Moreover, these methods struggle with scalability due to the scarcity of datasets with high-quality explanatory annotations essential for training models that provide more than binary "fake" or "real" labels.

A common challenge across both methodological directions in MFND is the heavy reliance on supervised training data. Building large-scale multimodal datasets with thousands of annotated samples is costly and labor-intensive, as it demands expert annotation to ensure accuracy and reliability.

To address these challenges, we propose a novel MLLM-guided active learning framework that reduces data requirements while sustaining high detection accuracy and interpretability. The approach leverages uncertainty-based sampling guided by MLLM reasoning to prioritize the most informative training samples, ensuring efficient annotation and improved model performance.

Further, to balance explainability with computational efficiency, the framework introduces a secondary lightweight classifier that learns to approximate MLLM reasoning patterns on selected samples. This dual-component design enables fast inference while preserving access to detailed, evidence-based explanations when necessary.

Preliminary experiments show that our method achieves performance comparable to fully supervised models while using less than 50% of the training data, reaching 90% accuracy on standard benchmarks. Moreover, the secondary classifier enhances overall accuracy by exploiting learned reasoning, making the framework both scalable and practical for real-world deployment.

The contributions of this work can be summarized as follows:

1. We propose a sequential hybrid strategy that intelligently combines entropy-based uncertainty sampling with MLLM-guided disagreement sampling.
2. We introduce a second-stage lightweight classifier that uses MLLM reasoning to improve performance on hard samples.
3. Our method demonstrates superior sample efficiency and robust performance compared to standard baselines using 100% of training data.
4. We show that our framework generalizes effectively across datasets, maintaining strong performance even when applied to new, unseen multimodal news corpora.

2. Related Literature

The primary objective in Multimodal Fake News Detection (MFND) is to develop models that can accurately verify the consistency and veracity of an image-caption pair. The prominent benchmark for this task is the **NewsClippings**[8] dataset, which consists of real-world news articles where fake samples are synthetically generated by pairing a genuine image from one article with a caption from another, creating a subtle but malicious semantic mismatch.

LLMs in Multimodal Fake News Detection Early approaches have leveraged Vision-Language Models (VLMs), particularly based on **CLIP** [13], to solve this classification task. Models such as **CCN** [1], **Reddot** [10], and **FRAUD-Net** [11] have built upon this, which project images and query captions into a shared embedding space. These methods introduce mechanisms like evidence fusion, cross-modal attention, and domain-aware classifiers to improve alignment verification. However, their black-box nature limits interpretability. To address this explainability gap, a second line of research has focused on MLLMs. For instance, **SNIFFER** [12] is a two-stage fine-tuned **InstructBLIP** [5] model, which is trained on **GPT-4** [3] generated data to both detect fake news and generate explanatory statements. While effective at providing reasoning, such models introduce significant practical challenges. The training and inference of these large models demand substantial computational resources, and they rely on a large scale of annotated reasoning data, which is both costly and labor-intensive to create.

Active Learning A machine learning paradigm that reduces labeling cost by selecting the most informative samples under limited budgets [14, 15]. Traditional query strategies rely on model uncertainty or data diversity [7]. Uncertainty-based methods, such as **Least Confidence**, **Margin**, and **Entropy** [6], label samples where the model is most confused, but face a "cold-start" issue [4], as models trained on tiny initial sets provide unreliable uncertainty estimates. **MHPL (Minimum Happy Points Learning)** [18] addresses this in source-free domain adaptation by selecting neighbor-chaotic, diverse, and source-dissimilar samples, jointly modeling uncertainty, diversity, and exploitation. This consistently outperforms standard AL baselines, highlighting the benefit of going beyond classic heuristics.

LLMs in the Active Learning Loop Recent work explores using pre-trained LLMs to enhance AL query strategies [19]. **ActiveLLM** [4], for example, employs an instruction-tuned LLM to select samples from unlabeled pools, mitigating cold-start and outperforming traditional methods. However, its internal selection logic remains opaque, making it a "black box." In contrast, we introduce a hybrid AL strategy for MFND that leverages LLM-guided disagreement sampling (Section 4.2) and a Reasoning-Aware Classifier to distill MLLM reasoning for the hardest cases. This cascaded approach balances efficiency, accuracy, and interpretability.

3. Problem Statement

Multimodal fake news detection (MFND) is crucial because misinformation often spreads through image–text pairs, where genuine images are paired with misleading or unrelated text. This out-of-context pairing exploits the credibility of multimedia, making automated detection challenging. Formally, a news item is represented as a query pair (I_q, T_q), with I_q the query image and T_q the query text. The task is to predict its veracity label $y \in \text{true}, \text{fake}$. Optionally, external evidences may be available: visual $I_k^e k = 1^{N_e}$ and textual $T_k^e k = 1^{N_e}$, where N_e is the number of evidences. The goal is to learn a model f_θ that, given (I_q, T_q) and (if available) $I_k^e, T_{k\,k=1}^{e\,N_e}$, outputs $\hat{y}$ at inference.

Key challenges in this setting include: (i) the limited interpretability of vision-language models, which hinders real-world adoption; (ii) the high computational and data requirements for LLM-based approaches that can address interpretability; and (iii) the scarcity of large, high-quality annotated datasets for multimodal misinformation.

We therefore seek a data-efficient and interpretable MFND framework that reduces annotation cost via LLM-guided hybrid active learning and improves performance on hard cases through a reasoning-aware classifier.

4. Methodology

Our proposed framework, Figure 1, LLM Guided Active Learning for Multimodal Fake News Detection (HybridNet), 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). The methodology is structured into three core stages designed to strategically minimize data annotation requirements while maximizing detection accuracy and model interpretability. First, a Vision-Language Model (VLM) particularly CLIP [13] based network that leverages the multimodal feature alignment in shared latent space for multimodal fake news detection; Second, a novel hybrid active learning strategy that leverages the reasoning capabilities of a Multimodal Large Language Model (MLLM) to strategically select the most informative samples for annotation; and Third, finally we propose a reasoning-aware classifier, a lightweight module trained on difficult samples to distill MLLM-generated reasoning patterns. This multi-stage approach significantly reduces the dependency on large-scale labeled datasets while maintaining high detection accuracy and enhancing model interpretability for hard samples.

4.1. Base-Network for MFND

In our implementation, we employ a CLIP based network, FRAUD-Net [11], as our base network, BaseNet (f_θ). This choice is motivated by its proven effectiveness and its sophisticated architecture, which fuses the information from the primary image-text pair with external multimodal evidence using transformer-based attention mechanisms. The state-of-the-art performance by our base architecture is 91.1%, which is obtained by exhausting all the data points during its training, while our proposed framework demonstrates that high accuracy (90%) can be attained with less than half of the total data.

Algorithm 1 Hybrid Active Learning with LLM-Guided Disagreement

Require: Unlabeled pool D_U; Annotation budget B; Selection fraction α; Phase transition iter. k; Pre-trained MLLM M; Annotation Oracle $\mathcal{O}$.
Ensure: Final BaseNet parameters θ.
Initialize:
$b \leftarrow \lfloor \alpha \cdot |D_U| \rfloor$ ▷ Set fixed batch size per iteration
$D_{seed} \leftarrow \text{RandomSample}(D_U, \text{size})$
$D_L \leftarrow \mathcal{O}(D_{seed})$ ▷ Annotate the initial seed set
$D_U \leftarrow D_U \setminus D_{seed}$
$i \leftarrow 1$ ▷ Initialize iteration counter

while $|D_L| < B$ **do**
  Train BaseNet $f_{\theta^{(i-1)}}$ on the current labeled set D_L.
  if $i \leq k$ **then** ▷ **Phase 1: Uncertainty Sampling**
    $D_{\text{select}} \leftarrow \underset{x_j \in D_U}{\arg\text{top-k}}\ b\, H(f_{\theta^{(i-1)}}(x_j))$
  else ▷ **Phase 2: LLM-Guided Disagreement**
    $D_{\text{disagree}} \leftarrow \{x_j \in D_U \mid f_{\theta^{(i-1)}}(x_j) \neq M(x_j)\}$
    $D_{\text{select}} \leftarrow \underset{x_j \in D_{\text{disagree}}}{\arg\text{top-k}}\ b\, LowConf(f_{\theta^{(i-1)}}(x_j))$
  end if
  Update:
  $D_L \leftarrow D_L \cup \mathcal{O}(D_{\text{select}})$ ▷ Annotate batch and add
  $D_U \leftarrow D_U \setminus D_{\text{select}}$ ▷ Remove from unlabeled pool
  $i \leftarrow i + 1$
end while
return Final trained model parameters $\theta^{(i-1)}$.

4.2. MLLM Guided Active Learning

To achieve high data efficiency, we introduce a novel hybrid active learning strategy designed to intelligently curate the training set, as detailed in Algorithm 1. The process begins with a small, randomly sampled and annotated seed set, D_L, and a large pool of unlabeled data, D_U. Our objective is to iteratively augment D_L by selecting the most informative samples from D_U for annotation, continuing until the size of the labeled set reaches a predefined annotation budget, B. In each iteration, a fixed batch of b samples is selected, where b is determined by a selection fraction α of the initial size of the unlabeled pool.

Our hybrid strategy is temporally phased to balance computational cost with sampling efficacy. An iteration

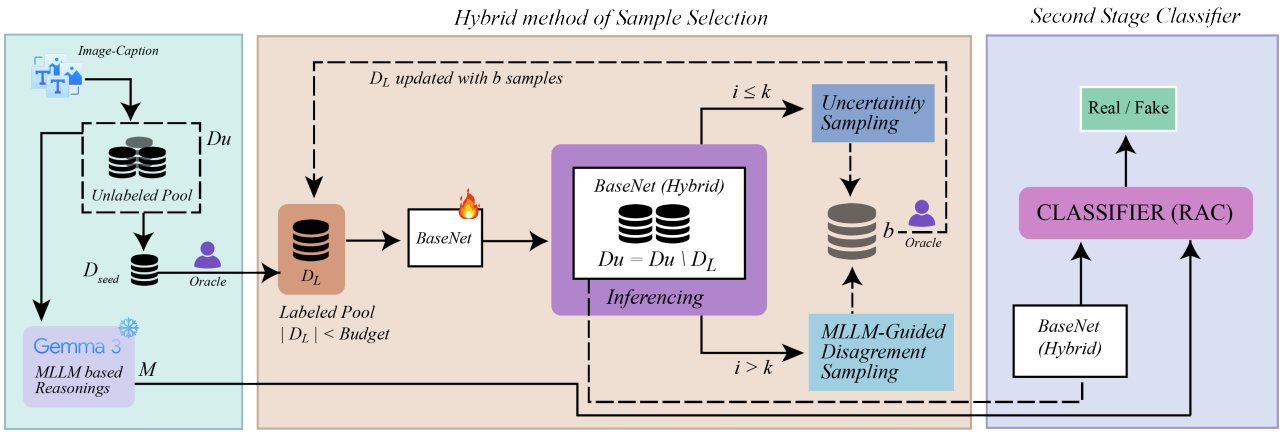

Figure 1. **HybridNet Framework.** A hybrid active learning loop trains BaseNet using uncertainty sampling (H) and MLLM (M) disagreement. The trained BaseNet is then combined with RAC for the final `Real/Fake` prediction.

counter, i, tracks the process. For the initial k iterations, when the BaseNet is nascent, we employ entropy-based uncertainty sampling. As the BaseNet matures and develops a more reliable decision boundary, we transition to a more sophisticated strategy based on the disagreement between the BaseNet and the MLLM prediction.

Phase 1: Initial Uncertainty-Based Seeding (Iterations $i \leq k$)
During the initial iterations, the BaseNet, denoted $f_{\theta^{(i-1)}}$, is trained on a small labeled set D_L and is thus prone to high uncertainty. At this stage, leveraging a computationally expensive MLLM is inefficient. We therefore opt for a classic and efficient uncertainty measure: Shannon entropy. The BaseNet produces a predictive probability $p_j = f_{\theta^{(i-1)}}(x_j)$ for each sample $x_j \in D_U$. The associated entropy is calculated as:

$$H(p_j) = -p_j \log_2(p_j) - (1 - p_j) \log_2(1 - p_j) \quad (1)$$

A batch of b samples with the highest entropy scores is then selected for annotation by the oracle.

Phase 2: LLM-Guided Disagreement Sampling (Iterations $i > k$)
Once the BaseNet has been trained on a sufficiently diverse set of samples, it has become a reasonably competent classifier. Therefore, in this phase, we first construct a disagreement set, D_{disagree}, containing all samples from the unlabeled pool where the BaseNet's predicted label differs from that of the MLLM:

$$D_{\text{disagree}} = \{x_j \in D_U \mid f_{\theta^{(i-1)}}(x_j) \neq M(x_j)\}. \quad (2)$$

The logic is that when both models agree, they are likely converging on well-understood, less ambiguous cases. In contrast, disagreement highlights instances where their representations diverge, suggesting that the sample is inherently more challenging or uncertain, and thus more informative for guiding the next round of training. From this set of conflicting predictions, we prioritize the samples on which the BaseNet is least confident. We select the top-b samples from D_{disagree} that exhibit the least confidence by BaseNet, effectively focusing our annotation budget on the most ambiguous and contentious cases. After each selection, the chosen batch is annotated and added to D_L, and the iteration counter is incremented. The BaseNet is then retrained on the newly expanded labeled set. This hybrid design first uses a cost-effective method to rapidly improve the nascent model, then deploys the MLLM to strategically identify and resolve the most challenging cases, ensuring both efficiency and the selection of high-quality, informative samples throughout the training cycle.

4.3. Reasoning-Aware Classifier

While our active learning strategy enhances the BaseNet's performance, certain challenging samples can still lead to misclassifications. To address this, we introduce, an expert model designed to enhance performance on "hard" cases. As specified by its implementation, it is a simple transformer based module that fuses BaseNet's internal features and MLLM's textual reasoning to further enhance BaseNet's performance.

Model Architecture The architecture of the Reasoning-Aware Classifier (RAC), denoted g_ϕ, is designed to effectively fuse the learned representations from the BaseNet i.e $\mathbf{z}_{f_\theta} \in \mathbb{R}^{D_{f_\theta}}$, with the explicit rationales provided by the pre-trained MLLM (M). These rationals correspond to dif-

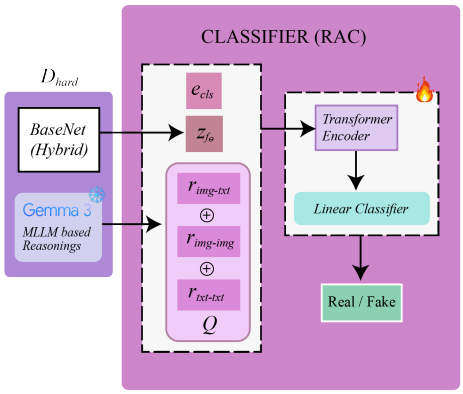

Figure 2. **Reasoning-Aware Classifier (RAC).** Illustration of how BaseNet features and reasoning embeddings are unified through projection, a classification token, and a Transformer encoder before final prediction.

ferent aspects of prompt based verification process whose embeddings $\mathbf{r}$ are generated by the $\mathbf{Q}$ qwen embedding model. Rational embeddings used are as follows:

1. **Image-Text Coherence** ($\mathbf{r}_{\text{img-txt}}$): Reasoning about the consistency between the query image and the query caption.
2. **Image-Image Similarity** ($\mathbf{r}_{\text{img-img}}$): Reasoning about the consistency between the query image to best external evidence image.
3. **Text-Text Verification** ($\mathbf{r}_{\text{txt-txt}}$): Reasoning about the factual veracity of the query caption to external evidence captions.

As illustrated in Figure 2, the model begins by unifying the BaseNet's feature embedding, $\mathbf{z}_{f_\theta}$ and the three distinct reasoning embeddings from the MLLM, $\mathbf{r}_{\text{img-txt}}$, $\mathbf{r}_{\text{img-img}}$, and $\mathbf{r}_{\text{txt-txt}}$ are projected into a common latent space, d_h, via separate linear transformations:

$$\mathbf{h}_j = \text{Proj}_j(\mathbf{k}_j), \quad \mathbf{k}_j \in \{\mathbf{z}_{f_\theta}, \mathbf{r}_{\text{img-txt}}, \mathbf{r}_{\text{img-img}}, \mathbf{r}_{\text{claim-ver}}\} \tag{3}$$

Subsequently, these projected embeddings are arranged into a sequence, and a learnable classification token, $\mathbf{e}_{\text{cls}} \in \mathbb{R}^{d_h}$, is added. The resulting sequence, $\mathbf{S} \in \mathbb{R}^{5\times d_h}$, encapsulates all available information:

$$\mathbf{S} = [\mathbf{e}_{\text{cls}}, \mathbf{h}_{f_\theta}, \mathbf{h}_{\text{img-txt}}, \mathbf{h}_{\text{img-img}}, \mathbf{h}_{\text{txt-txt}}] \tag{4}$$

This sequence is then processed by a Transformer Encoder, which leverages self-attention mechanisms to model the complex inter-dependencies between the BaseNet's features and the various MLLM reasoning aspects. The output embedding corresponding to the '[CLS]' token, which serves as an aggregated representation of the entire input, is then passed to a final MLP head to produce the output logit ℓ:

$$[\mathbf{s}'_0, \ldots, \mathbf{s}'_4] = \text{TransformerEncoder}(\mathbf{S}), \quad \ell = g_{\text{cls}}(\mathbf{s}'_0) \tag{5}$$

Training Objective The RAC is trained exclusively on a set of hard samples, $\mathcal{D}_{\text{hard}} = \{(x_j, y_j)\}_{j=1}^{M}$, where $y_j \in \{0, 1\}$ is the ground-truth label. These samples are curated by identifying instances where the BaseNet trained using our hybrid method either misclassifies or predicts with low confidence on the remaining pool $\boldsymbol{D_U}$ of the training data. The parameters ϕ of the RAC are optimized by minimizing the binary cross-entropy (BCE) loss. Given the final prediction $\hat{y}_j = \sigma(\ell_j)$, where σ is the sigmoid function, the training objective is to find the optimal parameters ϕ^*:

$$\phi^* = \arg\min_{\phi} \frac{1}{M} \sum_{(x_j, y_j)\in\mathcal{D}_{\text{hard}}} \mathcal{L}_{\text{BCE}}(y_j, \sigma(g_\phi(x_j))) \tag{6}$$

where $\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -[y \log(\hat{y}) + (1-y)\log(1-\hat{y})]$. The training process uses the Adam optimizer and incorporates an early stopping mechanism based on validation accuracy.

5. Experimental Evaluation

In this section, we conduct a series of experiments to rigorously evaluate the proposed framework. Based on our evaluation, we try to address below research questions:

- RQ1: *Can data-efficient training through active learning improve the effectiveness of multimodal fake news detection (MFND)?*
- RQ2: *Does our LLM-guided active learning strategy offer superior data efficiency compared to traditional baselines?*
- RQ3: *How effectively does the second-stage Reasoning-Aware Classifier (RAC) improve detection accuracy on challenging samples?*
- RQ4: *How do the choice of MLLM and prompt design influence reasoning quality and framework performance?*
- RQ5: *How well does our proposed framework generalize across different datasets and domains for multimodal fake news detection?*

We structure our experiments to first validate the active learning component, followed by an in-depth analysis and ablation of the RAC.

Dataset A popular benchmark for multimodal fake news detection, the **NewsClippings** [8] dataset, is used for all experiments. It contains real-world image–caption pairs, where fake samples are created by pairing images and captions 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 and ablations, we use the **IFND**

[16] dataset. While the full dataset contains more samples, we utilize a subset of 18k with predefined train–test splits for our experiments. Together with NewsClippings, these datasets provide a strong foundation for benchmarking and validating our framework.

Implementation Details We use FRAUD-Net with its official implementation as our BaseNet, built on a frozen CLIP [13] ViT-L/14 backbone. Active learning starts with a randomly sampled 10% seed set, fixed across all experiments. Performance is evaluated at annotation budgets of **20%**, **30%**, and **40%**. In our Hybrid strategy, the phase transition from uncertainty sampling to LLM-guided disagreement occurs after the sixth iteration ($k = 6$); We also tested multiple values of k and found that delaying the transition to later iterations generally results in greater performance. For cross-dataset generalization on IFND [16], we fine-tune the BaseNet with 1,000 samples and test on the remaining 17,000; the Hybrid strategy follows the same setup, with finetuning performed at the same annotation budgets. We adopt **Gemma** [17] as the guiding MLLM for its strong reasoning ability and consistent explanatory quality compared to Phi [2], Xgen [9], and Florence. RAC reasoning embeddings are generated using the pre-trained Qwen [20] model $\boldsymbol{Q}$. The RAC itself is a Transformer with hidden size 512, 2 attention heads, dropout 0.3, and is trained with Adam (1×10^{-4}) and early stopping (patience 5) based on validation accuracy.

5.1. Data-efficient training

Our first set of experiments evaluated whether active learning can improve data efficiency for MFND. We conducted these experiments on the **BaseNet**, since it demonstrates strong performance under full-data training, making it a reliable reference point for upper-bound comparisons. As shown in Figure 3, the entropy-based active learning curve rises more sharply than random sampling in the early stages, indicating that strategically chosen samples accelerate model learning. With only 20% of labeled data, active learning already outperforms random sampling by nearly 3 points, and by 50% it reaches 89.5% accuracy, closely approaching the fully supervised upper bound of 91.0%. In contrast, random sampling lags behind at the same budget, yielding only 87.8%. These results confirm that active learning saturates more quickly and that carefully selected instances provide disproportionate training value, establishing data-efficient training as both feasible and highly effective for MFND.

5.2. LLM-guided active learning

After demonstrating the benefits of data-efficient training, we evaluated whether LLM-guided selection surpasses conventional active learning heuristics (Table 1). At 10% labeling, all methods perform similarly ($\approx 83.9\%$), but differences widen at higher budgets: random sampling reaches 86.8% at 30%, entropy-based selection 88.5%, and MHPL [18] 87.4% at 40%, showing that diversity-aware strategies help but lag behind our hybrid approach.

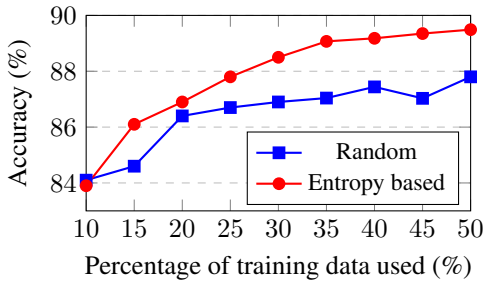

Figure 3. Learning curves showing NewsClippings test accuracy for Random sampling vs Entropy-based active learning at different percentages of labeled training data.

The hybrid method consistently surpasses all baselines, gaining 0.4–0.7 points after transitioning to LLM-guided disagreement at iteration $k = 6$ (35%). At 40% labeling, it achieves 89.6% versus 87.4% for random/MHPL and 89.2% for entropy. Adding RAC further boosts accuracy to 90.2%, demonstrating the benefits of combining reasoning-aware supervision with hybrid sampling. These results show LLM guidance effectively targets contextually challenging samples missed by traditional uncertainty or diversity heuristics.

For context, SNIFFER [12] reaches 88.4% but requires full-data training with higher computational cost, whereas our method achieves superior performance with less than half the labeled data, highlighting the efficiency gains of LLM-guided active learning. Performance gains saturate near the upper bound ($\approx 91\%$), indicating diminishing returns at larger budgets.

Strategy	**20%**	**30%**	**40%**
Random Sampling	86.2	86.8	87.4
MHPL	85.1	86.6	87.4
Uncertainty (Entropy)	86.9	88.5	89.2
Hybrid (Ours)	86.9	88.5	**89.6**
Hybrid + RAC (Ours)	**87**	**89.2**	**90.2**

Table 1. Test accuracy (%) of the BaseNet trained on NewsClippings [8] dataset with different active learning strategies at varying data budgets. Our Hybrid approach consistently outperforms other data-efficient baselines and rapidly approaches the fully supervised performance.

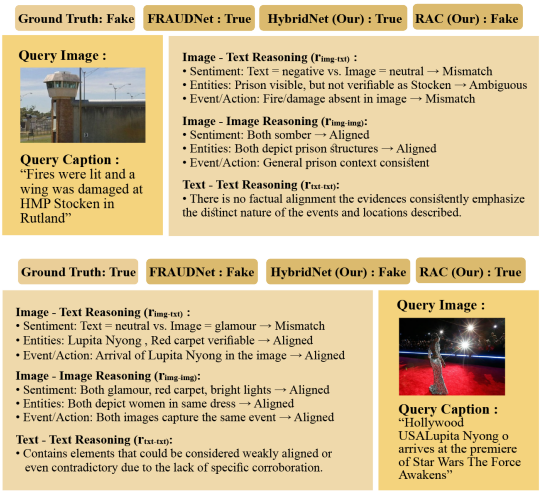

Figure 4. **Qualitative Examples of RAC.** Challenging NewsClippings samples where FRAUDNet and HybridNet fail, but RAC corrects the prediction using reasoning cues.

5.3. Efficacy of Reasoning-Aware Classifier (RAC) on challenging samples

The previous subsection showed that data-efficient training and hybrid active learning allow BaseNet to approach fully supervised performance using only a fraction of the labeled data. Nevertheless, some *hard samples* remain consistently misclassified or assigned low confidence. Standard networks provide little insight into these failures, motivating the use of large language models (LLMs), which excel at generating task-specific reasoning.

Training LLMs directly for every use case is computationally prohibitive. Instead, we extract reasoning features from a pre-trained LLM and fuse them with BaseNet embeddings to train a lightweight **Reasoning-Aware Classifier (RAC)**. Trained on $\sim$4000 hard samples ($\mathcal{D}_{\text{hard}}$), RAC achieves 83.9% accuracy on NewsClippings (Table 2) using only 5% labeled data, surpassing the 5% random BaseNet (79.1%) and nearly matching FraudNet at 10% random (84.1%), underscoring the strong data efficiency of our reasoning-based approach.

At inference, RAC acts as a second-stage verifier on low-confidence BaseNet predictions. With a 0.8 confidence threshold, 63 uncertain samples were flagged: BaseNet alone achieved 58% accuracy, while RAC improved this to 70%. Applied on top of the best Hybrid BaseNet (40% labeled data), RAC provides an additional 0.6% gain, reaching 90.2% accuracy—close to the fully supervised upper bound of 91.1%—while using only 45% labeled data (40% + 5%). Qualitative examples in Figure 4 further highlight RAC's ability to resolve inconsistencies that mislead other models.

To analyze the role of different inputs, we conducted an ablation study (Table 2) with three RAC variants: (i) BaseNet features only, (ii) LLM reasoning embeddings only, and (iii) the full model combining both. While each feature set alone already matches or slightly surpasses Hybrid BaseNet, the **full model achieves the highest accuracy of 90.2%**, confirming the complementary nature of learned representations and structured reasoning.

Model Configuration	Overall Accuracy (%)
BaseNet (Hybrid @ 40%)	89.6
FraudNet (@ 10%)	84.1
RAC	83.9
BaseNet + RAC (BaseNet Features Only)	90.02
BaseNet + RAC (Reasoning Features Only)	90.08
BaseNet + RAC (Full Model)	**90.2**
FraudNet [11]	91.1
FraudNet + RAC	**91.5**

Table 2. Ablation study on the Reasoning-Aware Classifier (RAC). Combining BaseNet and reasoning features gives the best gains (90.2%)

Overall Performance Boost Table 2 shows that the cascaded model ('BaseNet + RAC') consistently improves over the Hybrid BaseNet, raising accuracy from 89.6% to 90.2% by leveraging both learned features and LLM reasoning. The ablation confirms that each component contributes: using only BaseNet features (90.02%) or only reasoning features (90.08%) already matches or slightly surpasses the standalone BaseNet, while their combination delivers the strongest gains. Notably, RAC trained with only 5% labeled data already achieves 83.9%, which is comparable to FraudNet trained on 10% data (84.1%) on the NewsClippings test set, highlighting its data efficiency. Finally, when RAC is applied on top of the fully trained FraudNet [11], it increases accuracy from 91.1% to 91.5%, demonstrating RAC's ability to correct challenging cases and push performance beyond the baseline upper bound.

5.4. Impact of MLLM Choice and Prompt Design

To address RQ4, we evaluated multiple MLLMs (Gemma, Phi, Xgen, Florence) within our reasoning pipeline, which applies three complementary checks:

1. **Query image vs. query text** and **query image vs. evidence image** for entity, event, and sentiment consistency.
2. **Query text vs. evidence texts** for factual verification and alignment.

Each component uses a dedicated prompt, and their outputs are combined into a final classification and explanation. For the hybrid method, this reasoning-based prediction is compared with BaseNet outputs, with mismatches

prioritized for labeling and RAC training.

We systematically compared MLLMs by reasoning quality and model size. **Gemma** [17] (12B) consistently produced the most precise and contextually accurate reasoning, correctly identifying mismatches and aligning with evidence. By contrast, **Xgen**, Florence, and smaller variants of **Phi** [2] often generated generic or incorrect explanations. Gemma's detailed reasoning made it the clear choice, and it was used in all subsequent hybrid AL and RAC experiments.

5.5. Cross-Data Generalization

We evaluated the robustness of our approach on the **IFND** dataset, a collection of real-world news image-text pairs from the Indian news domain, which has a distribution different from NewsClippings.

Testing the fully supervised **FraudNet** model in a zero-shot (Figure 5) setting yielded only 58.3% accuracy, highlighting its poor cross-domain generalization. Finetuning FraudNet on a small subset of 1,000 IFND training samples (leaving 17,000 for testing) substantially improved performance to 89.4% overall, showing that limited adaptation enables effective generalization.

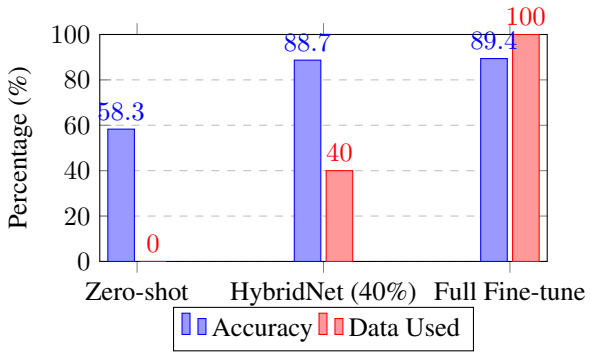

Figure 5. Cross-dataset performance comparison of Zero-shot BaseNet, HybridNet (40% data), and Full fine-tuning (100% data). Accuracy and training data usage are shown side by side.

To further reduce labeling cost, we applied data-efficient sampling strategies using the same hybrid active learning setup as in NewsClippings. Starting with a 10% random seed and incrementally adding 5% per iteration up to 40%, Table 3 shows that our hybrid method consistently outperforms Random, Entropy, and MHPL, achieving 88.9% accuracy at the 40% budget. Notably, with only ∼400 labeled samples (40% budget), our method already approaches the performance obtained by finetuning FraudNet on the full 1,000-sample training subset, demonstrating strong data efficiency and cross-domain generalizability. We did not further experiment with RAC on top of this setup due to the limited number of labeled samples available for training the reasoning-aware classifier.

Strategy	20%	30%	40%
Random	**78.2**	82.7	86.4
Entropy	73.6	80.3	87.9
MHPL [18]	75.4	83.8	86.8
Hybrid (Ours)	73.6	**84.1**	**88.9**

Table 3. Test accuracy (%) on IFND [16] dataset for various active learning strategies across different data budgets. Hybrid sampling consistently identifies informative samples, achieving competitive accuracy with fewer labeled instances.

5.6. Computational Requirements

All experiments were conducted on a single NVIDIA A6000 GPU (49 GB VRAM). The active learning phase is lightweight, with BaseNet training requiring only 3 GB of GPU memory. The first 8 iterations (batch size 16) complete within 4–5 hours, with training time scaling linearly as data increases. The only resource-intensive step is the one-time generation of MLLM reasonings for sample selection and RAC training, performed using Gemma-12B with vLLM (44 GB VRAM, 4 days for 70k NewsClippings samples).

Inference is highly efficient: the cascaded BaseNet+RAC model runs on 6 GB of VRAM, enabling deployment on standard GPUs. Compared to prior methods such as SNIFFER [12], which requires 4×A100 (40 GB) GPUs, our approach is far more practical. By leveraging open-source Gemma for reasoning, it also avoids the recurring API costs of proprietary models, providing a scalable, single-GPU solution with competitive performance.

6. Conclusion

In this work, we proposed a novel MLLM-guided hybrid active learning framework for multimodal fake news detection (MFND), effectively combining a state-of-the-art CLIP-based Vision-Language Model (FRAUD-Net) with the reasoning capabilities of a Multimodal Large Language Model. Our approach improves data efficiency and interpretability by selecting informative samples through uncertainty and disagreement-based sampling, and enhances detection on challenging cases via a lightweight Reasoning-Aware Classifier that leverages LLM explanations. Extensive experiments on the NewsClippings benchmark show accuracy comparable to fully supervised baselines using less than half the labeled data. Furthermore, cross-dataset evaluation on the IFND dataset demonstrates that our hybrid sampling strategy maintains strong performance and robust generalization, achieving near full-data finetuning accuracy using

only a fraction of labeled samples. The proposed framework offers a scalable, efficient, and interpretable solution for real-world multimodal misinformation detection, enabling more trustworthy and resource-conscious fake news detection systems.

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

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 6, 8

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[4] Markus Bayer. Activellm: Large language model-based active learning for textual few-shot scenarios. In *Deep Learning in Textual Low-Data Regimes for Cybersecurity*, pages 89–112. Springer, 2025. 2

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 2

[6] Tianxu He, Shukui Zhang, Jie Xin, Pengpeng Zhao, Jian Wu, Xuefeng Xian, Chunhua Li, and Zhiming Cui. An active learning approach with uncertainty, representativeness, and diversity. *The Scientific World Journal*, 2014(1):827586, 2014. 2

[7] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 5879–5899, 2024. 2

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

[9] Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450*, 2023. 6

[10] 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. 1, 2

[11] 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, 3, 7

[12] 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, 8

[13] 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, 3, 6

[14] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 2

[15] Burr Settles. Active learning literature survey. 2009. 2

[16] Dilip Kumar Sharma and Sonal Garg. Ifnd: a benchmark dataset for fake news detection. *Complex & intelligent systems*, 9(3):2843–2863, 2023. 6, 8

[17] 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, 8

[18] Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20008–20018, 2023. 2, 6, 8

[19] Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, et al. From selection to generation: A survey of llm-based active learning. *arXiv preprint arXiv:2502.11767*, 2025. 2

[20] 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