Multi-level Consistency and Conflict Decoupling with Knowledge from MLLM Layers for Multimodal Fake News Detection

Anonymous ACL submission

Abstract

Verifying multimodal news requires understanding semantic relations between images and texts. Real news usually shows aligned multimodal semantics, while fake news often contains inconsistencies at different levels, such as mismatches in objects, contexts, and claims. However, existing methods usually treat consistency and inconsistency in a coarse and static way, limiting their ability to adaptively exploit multi-level consistency and conflict clues. Inspired by the distinct knowledge encoded across layers in large language models, we propose MLCCD, a Multi-Level Consistency and Conflict Decoupling framework that exploits knowledge from different layers of multimodal large language models (MLLMs). Specifically, to obtain reliable knowledge from each layer, we first use the MLLM's final decision representation to select visual and textual information relevant to the task, reducing irrelevant content. Then, to obtain discriminative knowledge across layers, we preserve visual and textual information with distinctive changes between adjacent layers, capturing layer-specific hierarchical knowledge. Finally, we decouple the extracted knowledge into consistency and conflict evidence, and adaptively aggregate them across layers with separate routers. Experiments on multiple datasets show that our method outperforms existing methods. The code is available at https://anonymous.4open.science/r/MLCCD-8822.

1 Introduction

Verifying multimodal news requires understanding semantic relations between images and texts (Cheema et al., 2023). Compared with unimodal news, multimodal news conveys information through both visual content and textual descriptions, making it more expressive but also more vulnerable to manipulation (Tufchi et al., 2023). Real news usually shows aligned multimodal semantics (Xue et al., 2021; Müller-Budack et al., 2020), while fake news often contains inconsistencies at different levels (He et al., 2025). For example, the objects mentioned in the text may mismatch those shown in the image (Qi et al., 2021), the visual context may be inconsistent with the described event (Ma et al., 2024), or the two modalities may appear compatible while jointly implying an unsupported claim (Qi et al., 2024). These clues are often subtle and sparse (Abdali et al., 2024), but they are crucial for identifying fake news. Therefore, multimodal fake news detection requires models to capture not only the overall relation between images and texts, but also subtle consistency and conflict clues across different semantic levels.

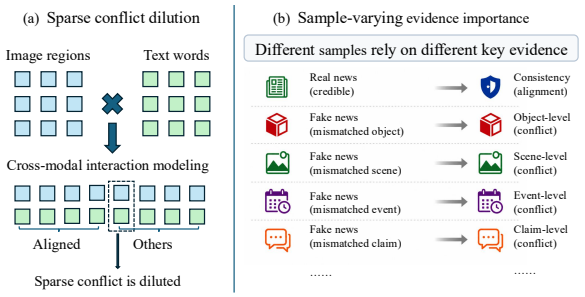

Figure 1: Motivating examples of sparse conflict dilution and sample-varying evidence importance in multimodal fake news detection.

Existing multimodal fake news detection methods have made promising progress by modeling relations between images and texts. Early studies usually learn joint multimodal representations to capture overall visual and textual semantics (Khattar et al., 2019; Chen et al., 2022), while recent methods further introduce image-text consistency modeling (Xue et al., 2021; Li et al., 2023), cross-modal attention (Wang et al., 2023, 2024), or external knowledge (Fu and Liu, 2023; Zhang et al., 2024) to improve multimodal reasoning. Nevertheless, these methods commonly model image-text consistency and conflict in a holistic and static manner. Such paradigms usually capture consistency

and conflict as coarse cross-modal signals, making it difficult to fully exploit subtle, sparse, and multi-level evidence that varies across different samples. Therefore, despite these advances, existing methods still face two important limitations.

First, sparse conflict clues are easily diluted. As illustrated in Figure 1(a), conflict clues in multimodal fake news usually appear in only a few image-text interactions. For example, the text may describe an object, action, or event detail that conflicts with what is shown in the image, while most other visual and textual content still appears reasonable and related. If all information is aggregated together, these few critical conflicts can be diluted by a large amount of normal or weakly related content. As a result, the learned multimodal representation may be dominated by less informative information, making sparse conflict clues difficult to preserve. Therefore, it is necessary to preserve task relevant and discriminative knowledge before constructing consistency and conflict evidence.

Second, consistency and conflict evidence contribute unevenly across samples. As illustrated in Figure 1(b), multimodal fake news verification depends on different evidence patterns for different samples. For real news, image and text are expected to support each other, so consistency evidence should usually play a more important role. For fake news, however, the decisive evidence may correspond to different conflict types. Some fake samples involve conflicts in local objects or visual details, while others rely on inconsistencies in event context, semantic relations, or textual claims. Therefore, the contribution of consistency and conflict evidence is not fixed, and different levels of conflict evidence may also contribute unevenly across fake samples. Fixed fusion strategies are difficult to adapt to such evidence variations, making it necessary to aggregate consistency and conflict evidence in a sample adaptive manner.

To address these challenges, we turn to the internal layer representations of MLLMs. Recent studies (Meng et al., 2022; Chuang et al., 2024) indicate that different LLM layers encode varied knowledge, from grammatical and surface level understanding in lower layers to semantic abstraction and reasoning capabilities in higher layers. Motivated by this, we propose MLCCD, a Multi-Level Consistency and Conflict Decoupling framework that exploits multi-level knowledge from MLLMs. Specifically, we first use the MLLM's final decision representation as a query to select task relevant visual and textual information, reducing the dilution of sparse conflict clues. Then, we measure changes between adjacent layers to preserve information with distinctive evolution, capturing discriminative knowledge across layers. Finally, we decouple the extracted representations into layer wise consistency and conflict evidence, and aggregate the two streams separately to adaptively emphasize the most informative levels for each sample.

Our main contributions are summarized as follows:

- To the best of our knowledge, this is the first work to exploit layer-wise knowledge from different MLLM layers for multimodal fake news detection. We show that such layer knowledge provides useful clues for modeling multi-level consistency and conflict evidence.
- We propose MLCCD, a Multi-Level Consistency and Conflict Decoupling framework. It reduces the dilution of sparse conflict clues through task relevant and discriminative knowledge extraction, and addresses uneven evidence contributions through consistency and conflict decoupling with adaptive aggregation.
- Extensive experiments on multiple datasets demonstrate that MLCCD outperforms existing methods, and further analyses validate the effectiveness of each component.

2 Related Work

2.1 Multimodal Fake News Detection

Multimodal fake news detection aims to verify news credibility by jointly considering textual content, visual information, and their semantic relations (Giachanou et al., 2020). Since fake news often conveys misleading intent through the mismatch between text and image (Xue et al., 2021), existing studies have moved beyond simple multimodal feature fusion (Khattar et al., 2019) toward cross modal alignment (Wang et al., 2023), consistency modeling (Xue et al., 2021), fine grained relation modeling (Liu et al., 2025), and knowledge enhanced verification (Gao et al., 2024). Recent methods further employ multimodal large language models to support explanation generation (Qi et al., 2024), commonsense reasoning (He et al., 2026), and evidence based judgment (Yan et al., 2025). However, most of these approaches still rely on

global representations, external knowledge, or generated rationales, while the internal layer knowledge of MLLMs remains largely underexplored. This limits the ability to understand and exploit how multimodal evidence is progressively encoded and transformed within MLLMs for fake news detection.

2.2 Multimodal Large Language Models

Multimodal large language models extend large language models to multimodal perception and reasoning by connecting visual encoders with language models through modality alignment modules (Bai et al., 2025; Grattafiori et al., 2024). Benefiting from large scale vision language pretraining (Chen et al., 2023) and multimodal instruction tuning (Zhao et al., 2023), they have shown strong capabilities in visual question answering (Davis et al., 2025), image captioning (Sarto et al., 2025), visual grounding (Xiao et al., 2025), instruction following (Huang et al., 2025), and complex multimodal understanding (Yin et al., 2024). Beyond final generated responses, MLLMs encode rich internal knowledge in layer wise hidden representations (Skean et al., 2025). Prior studies (Meng et al., 2022; Chuang et al., 2024) show that different layers capture different types of information, from surface level lexical and syntactic patterns to more abstract semantic and task related knowledge. This suggests that model knowledge is distributed across layers rather than concentrated only in the final representation. Recent studies (Lin et al., 2025; Li et al., 2026; Chen et al., 2025) further show that multimodal information also has layer dependent characteristics, as visual representations from different layers contribute differently to multimodal understanding and multi layer visual feature fusion can improve generalization. Therefore, layer wise hidden states provide a useful perspective for analyzing how multimodal knowledge is encoded and transformed within MLLMs.

3 Preliminaries

3.1 Task Definition

For the i-th multimodal news post, let X_i^t denote its textual content, X_i^v denote its visual content, and $y_i \in \{0, 1\}$ denote its ground truth label, where $y_i = 1$ indicates fake news and $y_i = 0$ indicates real news. Given N training examples, the dataset is denoted as $\mathcal{D} = \{(X_i^t, X_i^v, y_i)\}_{i=1}^N$. The goal of multimodal fake news detection is to learn a prediction function f_θ that estimates the fake news probability $\hat{y}_i = f_\theta(X_i^t, X_i^v)$ for each post.

3.2 MLLM Layer Representations

We use a LoRA tuned MLLM as the feature backbone. After tuning, the backbone is frozen, and the proposed evidence network operates on its hidden states. For each input post, the backbone produces layer wise token representations $\mathcal{H} = \{H^1, H^2, \ldots, H^L\}$. For layer ℓ, the hidden sequence $H^\ell = \{h_i^\ell\}_{i=1}^N$ consists of N token representations, where L is the number of MLLM layers and h_i^ℓ is the hidden state of the i-th token in the input multimodal sequence at layer ℓ. We use two binary modality masks, $M^t \in \{0, 1\}^N$ and $M^v \in \{0, 1\}^N$, to distinguish textual and visual tokens. Specifically, $M_i^t = 1$ indicates that position i belongs to the text modality, and $M_i^v = 1$ indicates that position i belongs to the image modality. Viewing binary masks as sets of selected positions, the valid token mask is defined as their union, $M = M^t \vee M^v$.

4 Methodology

4.1 Overview

MLCCD detects multimodal fake news by transforming the internal layer representations of an MLLM into structured verification evidence. Given a news post with an image and text, the MLLM encodes both modalities into a unified hidden sequence, where textual and visual elements are treated as tokens. Since decisive fake news clues are often sparse and may appear at different semantic depths, MLCCD progressively constructs evidence from token selection to layer knowledge extraction, and then performs fine-grained consistency and conflict decoupling across MLLM layers.

As shown in Figure 2, MLCCD consists of three modules. *Task Relevant Token Selection* uses the final decision representation as a query to select visual and textual tokens related to the verification decision. *Discriminative Layer Knowledge Extraction* retains selected tokens with distinctive representation changes between adjacent layers, capturing layer specific knowledge from the MLLM hierarchy. *Adaptive Consistency and Conflict Routing* builds separate evidence streams for cross modal agreement and discrepancy, and adaptively aggregates them across layers with independent routers.

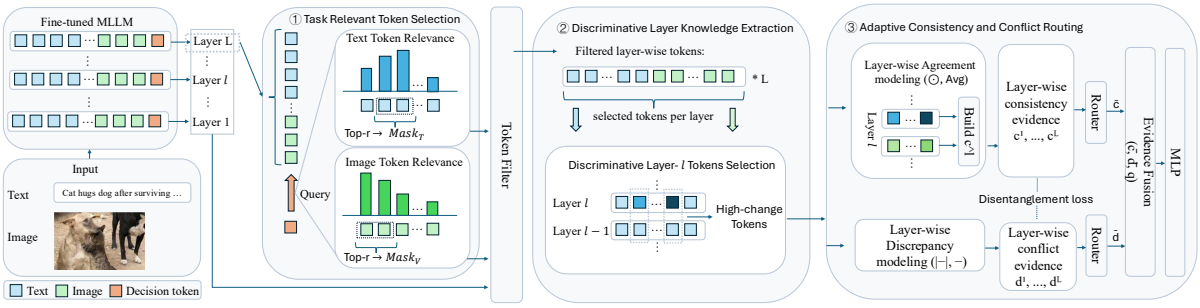

Figure 2: Overview of MLCCD. The frozen LoRA tuned MLLM provides layer wise hidden states, and the proposed evidence network selects task relevant tokens, extracts discriminative layer knowledge, and adaptively routes consistency and conflict evidence for fake news prediction.

4.2 Task Relevant Token Selection

This module constructs a compact evidence pool before layer knowledge extraction. Multimodal fake news clues are often localized in a small subset of textual or visual tokens, while many other tokens provide background information that is related to the post but less useful for verification. Directly aggregating all tokens may therefore dilute the decisive clues. To address this issue, we use the MLLM's decision context to measure token relevance and select informative textual and visual tokens separately.

In our prompt format, the MLLM is instructed to output the verification label as the first generated answer token. For an autoregressive MLLM, this answer is predicted from the hidden state of the final valid prompt token. Let $i^\star$ denote the position of this token and $q_0 = h_{i^\star}^L$ denote its hidden state in the last MLLM layer. This hidden state provides a compact decision context for the verification task, and we use it as a task query. To identify tokens that are most related to the pending verification answer, we compare this query with each final layer token in a shared evidence space:

$$s_i = \frac{(W_h h_i^L)^\top (W_q q_0)}{\sqrt{d}}. \tag{1}$$

Here, W_h and W_q are learnable projections, and d is the hidden dimension. For brevity, we denote the projected query $W_q q_0$ as q in subsequent modules. A larger s_i indicates that token i is more relevant to the verification decision and should be more likely to enter the evidence pool.

We then perform modality wise selection based on the relevance scores. For visual tokens, the candidate mask is $C^v = M^v$. For textual tokens, the decision token is removed from the candidate set, so $C^t = M^t \setminus \{i^\star\}$. For each modality $m \in \{t, v\}$, the selected token mask $\tilde{M}^m$ is obtained by top ratio selection:

$$\tilde{M}^m = \mathrm{TopR}(s, C^m, \rho_q^m), \quad m \in \{t, v\}, \tag{2}$$

where $\mathrm{TopR}(s, C^m, \rho_q^m)$ keeps the highest scored ρ_q^m proportion of tokens within the candidate mask C^m. The keep ratio ρ_q^m is modality dependent because the numbers of textual and visual tokens can differ substantially. This design allows each modality to preserve an appropriate amount of task relevant information and prevents the modality with more tokens from dominating the evidence pool. We also compute the normalized relevance distribution π_i^m over selected tokens:

$$\pi_i^m = \mathrm{softmax}_{i \in \tilde{M}^m}(s_i). \tag{3}$$

The relevance distribution π_i^m is later used as a reliability cue for layer wise evidence extraction.

4.3 Discriminative Layer Knowledge Extraction

After task relevant tokens are selected, this module further identifies which selected tokens carry discriminative layer knowledge. Different MLLM layers encode different levels of information, and a token whose representation changes strongly between adjacent layers is likely to contain layer specific evidence. Therefore, we use representation changes across layers to preserve tokens that are both task relevant and layer distinctive.

For token i at layer ℓ, the layer change score e_i^ℓ measures how much its representation changes from the previous layer. Since the first layer has no preceding layer, its score is set to zero as a boundary convention:

$$e_i^\ell = \begin{cases} 0, & \ell = 1, \\ \|h_i^\ell - h_i^{\ell-1}\|_2, & \ell > 1. \end{cases} \tag{4}$$

For each modality m and each layer ℓ, we standardize the layer change scores within the selected candidate tokens $\tilde{M}^m$ to obtain $e_{i,\text{std}}^{m,\ell}$, which measures the relative distinctiveness of token i within modality m at layer ℓ. The discriminative token mask $\bar{M}^{m,\ell}$ keeps the selected tokens with the largest standardized changes:

$$\bar{M}^{m,\ell} = \text{TopR}(e_{\text{std}}^{m,\ell}, \tilde{M}^m, \rho_e), \quad m \in \{t, v\}. \tag{5}$$

Here, ρ_e is the layer evidence keep ratio.

To integrate task relevance and layer distinctiveness, each retained token receives a weight $a_i^{m,\ell}$:

$$a_i^{m,\ell} = \text{softmax}_{i \in \bar{M}^{m,\ell}}(s_i + \gamma e_{i,\text{std}}^{m,\ell}), \tag{6}$$

The scalar γ is a learnable positive scale. The modality representation u_m^ℓ at layer ℓ is obtained by weighted pooling:

$$u_m^\ell = \sum_{i \in \bar{M}^{m,\ell}} a_i^{m,\ell} h_i^\ell, \quad m \in \{t, v\}. \tag{7}$$

To preserve modality identity and layer order, we project u_m^ℓ with a learnable projection W_f, and combine it with the layer position embedding p^ℓ and the modality embedding b_m:

$$\hat{z}_m^\ell = W_f u_m^\ell + p^\ell + b_m. \tag{8}$$

The resulting $\hat{z}_t^\ell$ and $\hat{z}_v^\ell$ are layer aware textual and visual representations that preserve task relevance and discriminative layer knowledge. In addition, the token weights are also used to summarize the reliability of the extracted evidence. Specifically, μ_m^ℓ measures retained decision relevance, and ν_m^ℓ measures retained layer knowledge strength:

$$\begin{aligned} \mu_m^\ell &= \sum_{i \in \bar{M}^{m,\ell}} a_i^{m,\ell} \pi_i^m, \\ \nu_m^\ell &= \sum_{i \in \bar{M}^{m,\ell}} a_i^{m,\ell} \sigma(e_{i,\text{std}}^{m,\ell}). \end{aligned} \tag{9}$$

These reliability scores are passed to the routing module together with the layer wise representations.

4.4 Adaptive Consistency and Conflict Routing

Given the extracted layer wise representations, MLCCD first applies a lightweight refinement before constructing consistency and conflict evidence. Specifically, a Transformer encoder E_{layer} is applied along the layer dimension to slightly contextualize each layer representation while keeping the layer wise outputs separate:

$$z_m^{1:L} = E_{\text{layer}}(\hat{z}_m^{1:L}). \tag{10}$$

Based on the refined representations z_t^ℓ and z_v^ℓ, MLCCD separates cross modal agreement from cross modal discrepancy at each layer. Consistency evidence reflects whether the visual and textual modalities support compatible semantics, while conflict evidence captures their divergent signals. The two evidence streams are then routed independently according to their layer wise reliability. In this way, decoupling is performed at multiple semantic levels instead of only on a single global multimodal representation.

For each layer ℓ, the consistency stream represents cross modal agreement with element wise interaction and averaged context:

$$g_c^\ell = \left[z_t^\ell \odot z_v^\ell; \frac{z_t^\ell + z_v^\ell}{2} \right]. \tag{11}$$

The conflict stream represents cross modal discrepancy with absolute and directional differences:

$$g_d^\ell = \left[|z_t^\ell - z_v^\ell|; z_t^\ell - z_v^\ell \right]. \tag{12}$$

The branch inputs from all layers are then transformed by feed forward builders ϕ_c and ϕ_d, followed by branch encoders E_c and E_d, producing layer wise consistency evidence $c^{1:L}$ and conflict evidence $d^{1:L}$:

$$\begin{aligned} c^{1:L} &= E_c(\phi_c(g_c^{1:L})), \\ d^{1:L} &= E_d(\phi_d(g_d^{1:L})). \end{aligned} \tag{13}$$

To provide the routers with evidence quality cues, we estimate the consistency strength κ_c^ℓ and conflict strength κ_d^ℓ with learnable scorers ω_c and ω_d:

$$\kappa_c^\ell = \sigma(\omega_c(c^\ell)), \quad \kappa_d^\ell = \sigma(\omega_d(d^\ell)). \tag{14}$$

Together with the modality relevance and layer knowledge strength from the previous module, these scores form the reliability vector r^ℓ:

$$r^\ell = [\mu_t^\ell; \mu_v^\ell; \nu_t^\ell; \nu_v^\ell; \kappa_c^\ell; \kappa_d^\ell]. \tag{15}$$

For branch $x \in \{c, d\}$, an independent router ψ_x computes the layer importance score η_x^ℓ from the

branch evidence x^ℓ, the decision query q, and the reliability vector r^ℓ:

$$\eta_x^\ell = \psi_x([x^\ell; q; r^\ell]). \tag{16}$$

The router keeps the top K layers for each branch:

$$\mathcal{K}_x = \mathrm{TopK}(\{\eta_x^\ell\}_{\ell=1}^L, K). \tag{17}$$

The routing weights are obtained by normalizing the retained layer scores:

$$\alpha_x^\ell = \frac{\exp(\eta_x^\ell)\mathbb{I}[\ell \in \mathcal{K}_x]}{\sum_{j\in\mathcal{K}_x}\exp(\eta_x^j)}. \tag{18}$$

The routed branch evidence $\bar{x}$ is computed as

$$\bar{x} = \sum_{\ell=1}^{L} \alpha_x^\ell x^\ell, \quad x \in \{c, d\}. \tag{19}$$

Thus, $\bar{c}$ and $\bar{d}$ denote the routed consistency evidence and routed conflict evidence, respectively. The routed reliability summary $\bar{r}$ averages the reliability vectors under the two routing distributions:

$$\bar{r} = \frac{1}{2}\left(\sum_{\ell=1}^{L}\alpha_c^\ell r^\ell + \sum_{\ell=1}^{L}\alpha_d^\ell r^\ell\right). \tag{20}$$

Finally, the classifier input z_{cls} combines routed evidence, evidence interaction, evidence difference, the decision query, and the routed reliability summary:

$$z_{\mathrm{cls}} = [\bar{c}; \bar{d}; \bar{c}\odot\bar{d}; |\bar{c}-\bar{d}|; q; \bar{r}]. \tag{21}$$

The fake news logit o and prediction probability $\hat{y}$ are produced by

$$o = \mathrm{MLP}(z_{\mathrm{cls}}), \quad \hat{y} = \sigma(o). \tag{22}$$

4.5 Training Objective

We train the evidence network with the standard binary cross entropy loss $\mathcal{L}_{\mathrm{cls}}$. To encourage the consistency and conflict streams to encode complementary evidence, we further use a disentanglement regularizer $\mathcal{L}_{\mathrm{dis}}$:

$$\mathcal{L}_{\mathrm{dis}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{L}\sum_{\ell=1}^{L}\cos(c_b^\ell, d_b^\ell)^2, \tag{23}$$

This term discourages the two streams from producing overly similar representations at each layer. With the disentanglement weight λ, the final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda\mathcal{L}_{\mathrm{dis}}. \tag{24}$$

During training, the LoRA tuned MLLM remains frozen, and only the three proposed modules and the classifier are optimized.

Dataset	Train	Dev	Test	Lang.
Weibo	7461	–	1929	Chinese
Pheme	1412	221	385	English
AMG	3268	501	883	English

Table 1: Dataset statistics.

5 Experiments

5.1 Datasets and Metrics

We evaluate MLCCD on three widely used multimodal fake news detection benchmarks: Weibo (Jin et al., 2017), Pheme (Zubiaga et al., 2017), and AMG (Guo et al., 2025). These datasets cover different languages, event sources, and data scales, enabling evaluation under diverse multimodal news verification scenarios. Table 1 summarizes the dataset statistics used in our experiments. The dash in the Dev column indicates that Weibo is evaluated with its original train/test split and does not provide a separate development set. For all datasets, we report Accuracy, macro F1, real class F1, and fake class F1 as the main evaluation metrics. The detailed preprocessing procedures and split settings are provided in Appendix A.

5.2 Baselines

We compare MLCCD with the baselines listed in Table 2. In the order of the table, CFFN (Li et al., 2023) is a fine grained fusion method that models consistency and inconsistency from word region relations. FKA-Owl (Liu et al., 2024) is an LVLM based fake news detector that augments large vision language models with forgery specific knowledge. MTS (Sun et al., 2025) models multimodal interactions through Taylor series expansion, and MIMoE-FND (Liu et al., 2025) uses mixture of experts to capture modality interactions. Among them, CFFN, MTS, and MIMoE-FND are conventional multimodal detectors, while FKA-Owl represents a large model based method. We further construct direct MLLM baselines by LoRA fine tuning Qwen-3-VL-2B-Ins (Bai et al., 2025), Qwen-3-VL-8B-Ins (Bai et al., 2025), Gemma-4-E2B-it (Google, 2026), and Gemma-4-E4B-it (Google, 2026) under the same data splits.

5.3 Implementation Details

We implement MLCCD with Qwen3-VL-8B-Instruct as the backbone MLLM. The backbone is first adapted with LoRA; during evidence network training, the LoRA tuned backbone is frozen

Method	Weibo				AMG				Pheme			
	Acc.	F1-Ma	F1-R	F1-F	Acc.	F1-Ma	F1-R	F1-F	Acc.	F1-Ma	F1-R	F1-F
CFFN (Li et al., 2023)	88.7	88.7	89.3	88.1	80.7	80.3	83.3	77.3	85.7	82.1	90.2	73.9
FKA-Owl (Liu et al., 2024)	88.7	88.7	88.3	89.1	85.5	85.1	87.6	82.6	87.0	85.1	90.4	79.8
MTS (Sun et al., 2025)	89.8	89.8	90.2	89.3	84.0	83.6	86.3	80.8	86.8	84.2	90.5	77.9
MIMoE-FND (Liu et al., 2025)	91.3	91.3	91.5	91.1	83.4	82.9	85.7	80.1	85.5	83.4	89.2	77.6
Qwen3-VL-2B + LoRA (Bai et al., 2025)	94.2	94.2	94.4	94.0	89.4	89.1	90.9	87.2	90.1	88.4	92.9	83.8
Qwen3-VL-8B + LoRA (Bai et al., 2025)	95.0	95.1	95.2	94.9	89.5	89.2	90.9	87.5	91.7	89.5	94.3	84.6
Gemma4-E2B + LoRA (Google, 2026)	91.8	91.9	92.0	91.7	84.9	84.6	86.8	82.4	91.2	89.4	93.8	85.0
Gemma4-E4B + LoRA (Google, 2026)	94.0	94.1	94.3	93.8	85.8	85.5	87.6	83.4	91.2	89.3	93.8	84.7
MLCCD	**95.3**	**95.3**	**95.3**	**95.2**	**90.4**	**90.1**	**91.8**	**88.3**	**93.2**	**91.8**	**95.3**	**88.2**
↑ Δ over 2nd	0.3	0.2	0.1	0.3	0.9	0.9	0.9	0.8	1.5	2.3	1.0	3.2

Table 2: Overall results on Weibo, AMG, and Pheme. All values are percentages. Best results are in bold, and second best results are underlined. F1-R and F1-F denote real class F1 and fake class F1. ↑ Δ over 2nd denotes the absolute point gain of MLCCD over the second best result in each column.

and only the proposed modules and the classifier are optimized. Each sample is formatted as one image and one text prompt. We train the evidence network with AdamW. For token selection, we set $\rho_e = 0.5$, keep at least one token for each non-empty top-ratio candidate set, fix $\rho_q^v = 0.3$, and set ρ_q^t to 0.3, 0.025, and 0.1 on Weibo, AMG, and Pheme, respectively, to account for different text lengths. Detailed hyperparameters and input processing settings are provided in Appendix B.

5.4 Overall Performance

Table 2 reports the overall results. MLCCD achieves the best accuracy and macro F1 on all three datasets, showing consistent gains over both prior multimodal detectors and direct LoRA tuned MLLM baselines. Compared with the strongest LoRA baseline, MLCCD improves accuracy by 0.3, 0.9, and 1.5 points on Weibo, AMG, and Pheme, respectively. The corresponding macro F1 gains are 0.2, 0.9, and 2.3 points.

The gains are modest on Weibo, where strong MLLM baselines already obtain high performance, and become more visible on AMG and Pheme. These results indicate that using internal MLLM layers as structured evidence can provide additional benefits beyond direct label supervised tuning. The advantage is larger when compared with existing fake news detection methods: MLCCD improves over the strongest baseline in this group by 4.0, 4.9, and 6.2 accuracy points on Weibo, AMG, and Pheme, respectively.

Class wise results further show that the gains are not concentrated in a single label. MLCCD improves fake news F1 consistently and also maintains strong real news F1, with the detailed precision, recall, and F1 scores reported in Appendix C.

5.5 Ablation Study

We conduct ablation studies on Pheme to examine the contribution of each component under the same training protocol. **w/o TRTS** removes Task Relevant Token Selection and uses all modality tokens as the candidate evidence pool. **w/o DLKE** removes Discriminative Layer Knowledge Extraction and does not select tokens according to layer transition scores. **w/o ACCR** removes Adaptive Consistency and Conflict Routing, and directly averages the layer aware textual and visual representations for prediction. **Shared Router** keeps the consistency and conflict streams but uses one shared router for both branches. **w/o Dis.** removes the disentanglement regularizer $\mathcal{L}_{\text{dis}}$.

Variant	Acc.	F1-Ma	F1-R	F1-F
w/o TRTS	92.2	90.5	94.5	86.4
w/o DLKE	92.2	90.7	94.4	87.0
w/o ACCR	92.5	90.7	94.8	86.6
Shared Router	91.7	89.5	94.3	84.6
w/o Dis.	92.7	91.1	94.9	87.3
MLCCD	**93.2**	**91.8**	**95.3**	**88.2**

Table 3: Ablation results on Pheme. All values are percentages. Best results are in bold. F1-R and F1-F denote real class F1 and fake class F1.

Table 3 shows that all ablated variants have lower accuracy and macro F1 than the full model. Removing TRTS or DLKE reduces macro F1 by 1.3 and 1.1 points, respectively, indicating that both task relevant token selection and layer transition based evidence extraction contribute to reliable verification. Removing ACCR also lowers macro F1 by 1.1 points, suggesting that simple layer averaging is less effective than explicitly organizing consistency and conflict evidence. The shared router variant obtains the largest drop, especially on fake news F1, which indicates that consistency and conflict evi-

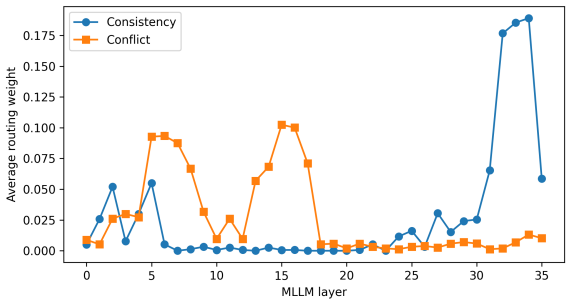

Figure 3: Average routing weights of the consistency and conflict branches across MLLM layers on Pheme.

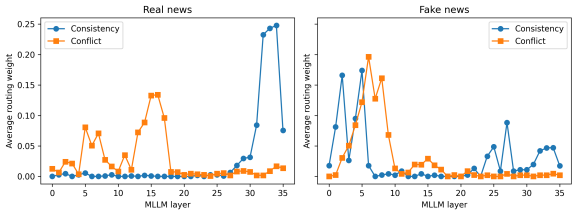

Figure 4: Routing weights grouped by real and fake news on Pheme.

dence benefit from separate layer routing. Removing $\mathcal{L}_{\text{dis}}$ causes a smaller but consistent degradation, showing that the disentanglement regularizer further improves the separation of the two evidence streams.

5.6 Layer Wise Evidence Analysis

We further analyze the trained model on the Pheme test set to examine how MLCCD uses the internal MLLM layers after training. For each sample, we record the routing distributions of the consistency and conflict branches, the overlap between their top routed layers, and the compactness of the selected evidence tokens. We focus on routing weights because they directly determine which layers contribute to the final prediction.

Figure 3 shows that the two evidence streams rely on different parts of the backbone. The consistency branch is concentrated in the upper layers: layers 28–35 receive 74.0% of its total routing weight, and the strongest region, layers 32–34, alone accounts for 55.1%. In contrast, the conflict branch mainly routes to middle layers, assigning 81.6% of its weight to layers 5–17, while only 5.2% is assigned to layers 28–35. This separation suggests that MLCCD does not simply duplicate the same layer evidence for both branches; instead, the two routers learn distinct depth preferences.

The separation is also clear at the sample level. Using the top four routed layers from each branch, 73.8% of samples have no overlap between the consistency and conflict branches, and the average top-K overlap is only 7.9%. The two routing distributions also have an average Jensen–Shannon divergence of 0.639 on the natural-log scale and an average cosine distance of 0.922. Meanwhile, the average layer-wise cosine similarity between the consistency and conflict evidence representations is close to zero (0.012). These results indicate that the decoupling is reflected not only in layer selection, but also in the evidence representation space.

Figure 4 further shows that the routing behavior changes with the label. For real news, the consistency branch is highly concentrated in the upper layers, with 96.2% of its weight on layers 28–35 and 72.4% on layers 32–34. For fake news, this upper-layer concentration is much weaker: layers 32–34 receive only 13.6% of the consistency weight, while the strongest peaks move to earlier and middle layers. The conflict branch mainly stays in lower-to-middle layers, but its detailed emphasis also changes: fake news places 84.3% of the conflict weight on layers 0–9, whereas real news assigns more conflict weight to layers 13–17. This pattern suggests that MLCCD does not apply a fixed depth preference to all inputs, but instead adapts the evidence levels used by each branch by shifting routing weights toward layers that provide more useful consistency or conflict cues for the current verification case.

6 Conclusion

In this paper, we propose MLCCD, a multi level consistency and conflict decoupling framework for multimodal fake news detection. Instead of relying only on the final representation or generated response of an MLLM, MLCCD transforms its layer wise hidden states into structured verification evidence. It first selects task relevant textual and visual tokens, then extracts discriminative layer knowledge from representation changes, and finally decouples cross modal evidence into consistency and conflict streams. With independent layer routing, the model can adaptively emphasize useful agreement and discrepancy cues for different samples. Experiments on three benchmarks demonstrate that MLCCD consistently outperforms prior multimodal fake news detectors and LoRA tuned MLLM baselines.

Limitations

The current framework depends on the quality of the MLLM hidden states. If the backbone fails to encode subtle visual manipulations or implicit factual relations, the evidence network may not be able to recover these missing cues. Moreover, ML-CCD is trained on benchmark datasets with specific event distributions, so its robustness under unseen topics, emerging events, or different social media platforms still needs further validation. Finally, the current detector does not retrieve external articles or real time evidence, which may limit its ability to verify claims that require fresh or specialized world knowledge beyond the backbone.

Ethics Statement

This work aims to support the detection of multimodal misinformation and should be used as an assistive technology rather than as a final arbiter of truth. We conduct experiments on existing benchmark datasets and do not collect new personal data. Since benchmark posts may still contain real social media content, researchers and practitioners should follow the licenses, privacy requirements, and data usage policies associated with each dataset. The predictions of a fake news detector can have social consequences: false positives may unfairly flag legitimate content, while false negatives may allow harmful misinformation to spread. Therefore, practical deployments should include human review, clear uncertainty reporting, appeal or correction mechanisms, and periodic evaluation across topics, languages, and communities. We also caution against using this technology for censorship, surveillance, or targeting individuals, and recommend documenting its intended scope and known limitations whenever it is released or deployed.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Gullal S Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A Bateman, and Ralph Ewerth. 2023. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in artificial intelligence*, 6:1125533.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Haoran Chen, Junyan Lin, Xinhao Chen, Yue Fan, Xin Jin, Hui Su, Jianfeng Dong, Jinlan Fu, and Xiaoyu Shen. 2025. Rethinking visual layer selection in multimodal llms. *arXiv e-prints*, pages arXiv–2504.

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*, pages 2897–2905.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.

Alexander Davis, Justin Parker, and Julian Perry. 2025. Image and video question answering with large language models: A comprehensive review.

Lifang Fu and Shuai Liu. 2023. Multimodal fake news detection incorporating external knowledge and user interaction feature. *Advances in Multimedia*, 2023(1):8836476.

Xingyu Gao, Xi Wang, Zhenyu Chen, Wei Zhou, and Steven CH Hoi. 2024. Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Transactions on Multimedia*, 26:8312–8322.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal fake news detection with textual, visual and semantic information. In *International Conference on Text, Speech, and Dialogue*, pages 30–38. Springer.

Google. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Accessed: 2026-05-26.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 228–236.

Dongxiao He, Yiqi Dong, Xiaobao Wang, Meng Ge, Carl Yang, Di Jin, and Witold Pedrycz. 2025. Unveiling fine-grained deceptive patterns in multi-modal fake news: An explainable neuro-symbolic framework with lvlms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2026. Factguard: Event-centric and commonsense-guided fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 363–371.

Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. 2025. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, 133(11):8151–8189.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

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*, pages 2915–2921.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18426–18434.

Jun Li, Yi Bin, Jie Zou, Jiwei Wei, Guoqing Wang, and Yang Yang. 2023. Cross-modal consistency learning with fine-grained fusion network for multimodal fake news detection. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pages 1–7.

Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, Bin Li, and Xiangyang Xue. 2026. Instruction-guided fusion of multi-layer visual features in large vision-language models. *Pattern Recognition*, 170:111932.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multi-layer visual feature fusion in multimodal llms: Methods, analysis, and best practices. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4156–4166.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. 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.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1212–1220.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. 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.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.

Jiahao Sun, Chen Chen, Chunyan Hou, Yike Wu, and Xiaojie Yuan. 2025. Multimodal taylor series network for misinformation detection. In *Proceedings of the ACM on Web Conference 2025*, pages 2540–2548.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. Fake news detection via multi-scale semantic alignment and cross-modal attention.

In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2406–2410.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5696–5704.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2025. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Kaiying Yan, Moyang Liu, Yukun Liu, Ruibo Fu, Zhengqi Wen, Jianhua Tao, and Xuefei Liu. 2025. Debunk and infer: Multimodal fake news detection via diffusion-generated evidence and llm reasoning. *arXiv preprint arXiv:2506.21557*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16777–16785.

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. 2023. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International conference on social informatics*, pages 109–123. Springer.

A Dataset Details

We evaluate MLCCD on three multimodal fake news benchmarks: Weibo (Jin et al., 2017), Pheme (Zubiaga et al., 2017), and AMG (Guo et al., 2025). We follow the standard split and preprocessing protocol of each dataset where available. For Weibo, the original train/test split is kept, and when a post contains multiple images, the image with the highest resolution is used as the visual input. Following common preprocessing practice (Chen et al., 2022; Lao et al., 2024), we remove samples with empty text, including 50 training samples and 66 test samples, as well as one test sample without a valid image. For Pheme, we adopt the same train/validation/test split as (Zheng et al., 2022). For AMG, we remove samples whose media files are not images, corrupted images, and images with abnormal resolution such as 1×1. Since AMG contains substantially longer texts, we truncate each AMG text to the first 1024 words during preprocessing. Table 1 reports the resulting statistics. For evaluation, we report Accuracy, macro F1, real class F1, and fake class F1.

B Implementation Details

For LoRA adaptation, we use rank 256, scaling factor 512, dropout 0.05, and apply LoRA to the query, key, value, output, gate, up, and down projection layers. For Qwen based models, the processor uses left padding and constrains the visual token budget to 4–256 tokens by setting `min_pixels` $= 4 \times 32^2$ and `max_pixels` $= 256 \times 32^2$. For Gemma based baselines, we use 280 visual soft tokens, which is the supported budget closest to our setting.

The evidence network follows a compact configuration. We set the hidden dimension to 1024 and the classifier fusion dimension to 512. Each lightweight Transformer encoder contains one block with 8 attention heads and dropout 0.1. For token selection, the layer evidence keep ratio is $\rho_e = 0.5$, and each top ratio operation keeps at least one token when the candidate set is non-empty. The image keep ratio is fixed as $\rho_q^v = 0.3$ across datasets. Since text lengths vary more substantially, we set the text keep ratio ρ_q^t to 0.3 on Weibo, 0.025 on AMG, and 0.1 on Pheme; the smaller value on AMG accounts for its much longer texts. During routing, each evidence stream keeps the top 4 layers. We train the evidence network for 5 epochs with AdamW, using learning rate 5×10^{-5}, weight decay 0.01, warmup ratio 0.1, batch size 24, gradient clipping 1.0, and DeepSpeed ZeRO stage 2 with bfloat16.

C Detailed Experimental Results

Tables 4, 5, and 6 report the full class wise results behind Table 2. All values are percentages. In contrast to the compact main table, these tables include precision, recall, and F1 for both fake and real news, which makes it easier to inspect whether a method improves one class at the expense of the

other.

Table 4 gives the detailed Weibo results. MLCCD obtains the best overall accuracy and F1 scores, while maintaining balanced fake and real class performance.

Method	Acc.	Fake News			Real News		
		Prec.	Rec.	F1	Prec.	Rec.	F1
CFFN	88.7	90.2	86.1	88.1	87.5	91.3	89.3
FKA-Owl	88.7	83.7	95.2	89.1	94.8	82.6	88.3
MTS	89.8	90.4	88.2	89.3	89.2	91.3	90.2
MIMoE-FND	91.3	90.0	92.3	91.1	92.6	90.4	91.5
Qwen3-VL-2B + LoRA	94.2	**94.4**	93.6	94.0	94.0	**94.8**	94.4
Qwen3-VL-8B + LoRA	95.0	**94.4**	95.4	94.9	95.6	94.7	95.2
Gemma4-E2B + LoRA	91.8	90.3	93.0	91.7	93.3	90.7	92.0
Gemma4-E4B + LoRA	94.0	94.3	93.4	93.8	93.8	94.7	94.3
MLCCD	**95.3**	93.3	**97.2**	**95.2**	**97.3**	93.5	**95.3**

Table 4: Detailed performance on Weibo.

Table 5 gives the detailed AMG results. The gains on AMG are visible in fake news precision and real news recall, indicating that the improvement is not limited to one class.

Method	Acc.	Fake News			Real News		
		Prec.	Rec.	F1	Prec.	Rec.	F1
CFFN	80.7	71.2	84.5	77.3	88.9	78.4	83.3
FKA-Owl	85.5	77.2	88.9	82.6	92.2	83.4	87.6
MTS	84.0	75.6	86.8	80.8	90.8	82.3	86.3
MIMoE-FND	83.4	74.6	86.5	80.1	90.5	81.3	85.7
Qwen3-VL-2B + LoRA	89.4	81.8	93.3	87.2	95.3	86.9	90.9
Qwen3-VL-8B + LoRA	89.5	80.9	**95.3**	87.5	**96.7**	85.8	90.9
Gemma4-E2B + LoRA	84.9	75.2	91.2	82.4	93.6	81.0	86.8
Gemma4-E4B + LoRA	85.8	76.3	92.1	83.4	94.3	81.9	87.6
MLCCD	**90.4**	**83.2**	94.2	**88.3**	96.0	**88.0**	**91.8**

Table 5: Detailed performance on AMG.

Table 6 gives the detailed Pheme results. MLCCD achieves the best fake news F1 and real news F1, which is consistent with the larger macro F1 gain reported in the main results.

Method	Acc.	Fake News			Real News		
		Prec.	Rec.	F1	Prec.	Rec.	F1
CFFN	85.7	79.6	69.0	73.9	87.8	92.6	90.2
FKA-Owl	87.0	73.3	**87.6**	79.8	**94.4**	86.8	90.4
MTS	86.8	76.3	79.6	77.9	91.4	89.7	90.5
MIMoE-FND	85.5	70.8	85.8	77.6	93.5	85.3	89.2
Qwen3-VL-2B + LoRA	90.1	81.0	86.7	83.8	94.3	91.5	92.9
Qwen3-VL-8B + LoRA	91.7	**92.6**	77.9	84.6	91.4	**97.4**	94.3
Gemma4-E2B + LoRA	91.2	85.0	85.0	85.0	93.8	93.8	93.8
Gemma4-E4B + LoRA	91.2	86.2	83.2	84.7	93.1	94.5	93.8
MLCCD	**93.2**	90.7	85.8	**88.2**	94.2	96.3	**95.3**

Table 6: Detailed performance on Pheme.

D Additional Layer Wise Evidence Analysis

This section provides the detailed statistics and plots used to support the layer wise analysis in Section 5.6. We focus on routing weights and representation separation, because these quantities directly describe how MLCCD selects layer evidence. The raw evidence-strength scores are learned reliability cues for the router rather than calibrated semantic conflict magnitudes, so we do not use them as direct evidence for label-wise conflict intensity.

Table 7 summarizes the numerical results used in the main text.

Quantity	Value
Samples / Accuracy	385 / 93.2
Cons. weight on layers 28–35	74.0
Cons. weight on layers 32–34	55.1
Conf. weight on layers 5–17	81.6
Conf. weight on layers 28–35	5.2
No-overlap samples	284/385
Mean top-4 overlap	7.9% * 4
Jensen–Shannon divergence	0.639
Routing cosine distance	0.922
Stream cosine similarity	0.012
Real cons. weight on layers 28–35	96.2
Real cons. weight on layers 32–34	72.4
Fake cons. weight on layers 32–34	13.6
Fake conf. weight on layers 0–9	84.3
Real conf. weight on layers 13–17	52.4

Table 7: Detailed evidence analysis statistics on Pheme. Values are percentages except the sample count, divergence, and cosine based metrics.

Figure 5 reports the representation-level separation between the two evidence streams. The cosine similarity remains close to zero across layers for both labels, indicating that the two streams are separated not only by their routing weights but also by their learned representations.

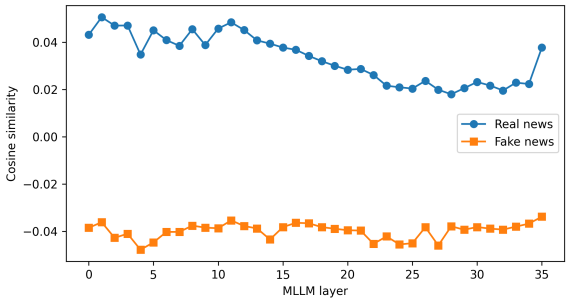

Figure 5: Layer-wise cosine similarity between consistency and conflict representations grouped by label.