DIVER: Dynamic Iterative Visual Evidence Reasoning for Multimodal Fake News Detection

Anonymous ACL submission

Abstract

Multimodal fake news detection is crucial for mitigating adversarial misinformation. Existing methods, relying on static fusion or LLMs, face computational redundancy and hallucination risks due to weak visual foundations. To address this, we propose DIVER (Dynamic Iterative Visual Evidence Reasoning), a framework grounded in a progressive, evidence-driven reasoning paradigm. DIVER first establishes a strong text-based baseline through language analysis, leveraging intra-modal consistency to filter unreliable or hallucinated claims. Only when textual evidence is insufficient does the framework introduce fine-grained visual evidence, where inter-modal alignment verification adaptively determines whether deeper visual inspection is necessary. For samples exhibiting significant inter-modal semantic discrepancies, DIVER selectively invokes fine-grained visual tools (e.g., OCR and Dense Captioning) to extract task-relevant evidence, which is iteratively aggregated via uncertainty-aware fusion to refine multimodal reasoning. Experiments on Weibo, Weibo21, and GossipCop demonstrate that DIVER outperforms the best baselines by 2.15% in accuracy on average, while reducing inference latency to 4.12 s. To facilitate reproducibility, the complete source code is available at an anonymous repository: https://anonymous.4open.science/r/DIVER-C4BE.

1 Introduction

The evolution of multimodal fake news has become increasingly adversarial, with fabricators pairing misleading text with contextually ambiguous imagery to complicate verification. While early studies focused on unimodal cues [3, 39], recent efforts have shifted toward multimodal fake news detection (MFND), where effective reasoning over inter-modal evidence is essential.

Recent MFND frameworks increasingly rely on large language models (LLMs) to enhance multimodal reasoning, such as INSIDE [34] and LIFE [31, 27]. Fundamentally, these approaches are static because they indiscriminately apply the same depth of visual processing to all inputs, failing to distinguish between obvious truths and subtle fabrications. Conversely, DIVER operates on a dynamic paradigm: it functions as an active decision-maker that regulates reasoning depth, triggering iterative visual evidence gathering only when high-level semantic alignment fails. [15, 1] Although leveraging LLMs for inference, these methods operate through single-pass interactions with global visual embeddings. This coarse representation fails to bridge the semantic gap between abstract claims and concrete visual facts. [26, 10, 16] Without explicit verification anchored in discrete visual evidence, LLMs are prone to hallucination when subtle visual details contradict textual narratives, leading to the validation of false claims based on unsupported contexts.

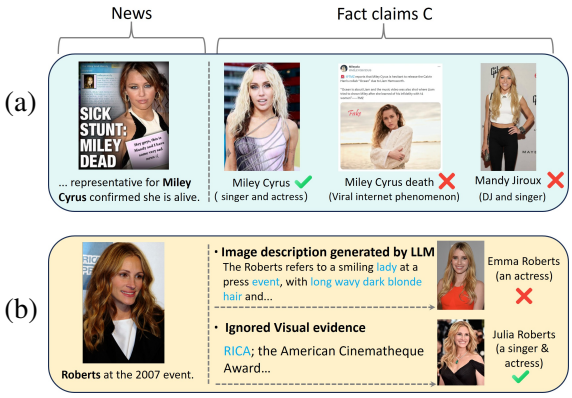

Figure 1: Two primary limitations of existing multimodal methods: (a) noise-induced factual ambiguity, and (b) visual hallucination due to weak grounding.

In contrast, traditional multimodal fusion approaches [40, 45] adopt static processing pipelines that indiscriminately incorporate deep visual features for all samples. While exhaustive visual forensics can improve robustness, it introduces

two inherent limitations. First, mandatory deep visual analysis incurs substantial computational overhead, which is redundant for samples that can be resolved through shallow semantic alignment. Second, forcibly injecting complex visual features when textual evidence is already sufficient may introduce irrelevant noise, potentially degrading detection accuracy. [8, 13, 11] These limitations reveal a fundamental mismatch between reasoning depth and computational efficiency in existing MFND systems.

To address this mismatch, MFND requires a dynamic mechanism that regulates reasoning depth based on inter-modal necessity. Specifically, visual forensics should be selectively activated only when semantic discrepancies between text and image indicate potential deception, rather than being uniformly applied to all inputs.

Based on this principle, we propose DIVER (Dynamic Iterative Visual Evidence Reasoning), which formulates multimodal fake news detection as a dynamic decision-making process. DIVER first establishes a strong text-based baseline to assess intra-modal consistency and eliminate unreliable claims. It then evaluates coarse inter-modal alignment and utilizes an inter-modal alignment score to quantify the semantic consistency between textual claims and visual representations. When this alignment score falls below a validation-selected threshold, the framework transitions from fast screening to deliberate visual reasoning. This process selectively invokes fine-grained forensic tools to extract explicit evidence for verification, reflecting a dual-stage reasoning structure analogous to System 1 and System 2 processes.

This alignment-guided gating strategy enables DIVER to effectively balance accuracy and efficiency. Across three benchmarks, DIVER achieves 95.8% accuracy on Weibo, 97.7% on Weibo21 and 91.6% on GossipCop, while reducing average inference latency to 4.12 s. In comparison, representative baselines such as LIFE and INSIDE incur higher latencies of 7.09 s and 4.97 s, respectively. By concentrating computational resources on samples exhibiting significant inter-modal inconsistencies, DIVER accelerates inference while mitigating hallucinations caused by irrelevant visual details.

Our contributions are summarized as follows:

- We formulate multimodal fake news detection as a dynamic decision-making process that prioritizes linguistic verification and selectively activates visual reasoning only when inter-modal discrepancies arise.
- We propose an alignment-guided gating mechanism that dynamically adjusts the granularity of analysis, enabling efficient filtering of low-risk samples while selectively invoking fine-grained visual forensics (e.g., OCR and Dense Captioning) for ambiguous cases.
- Extensive experiments across multiple benchmarks demonstrate that DIVER consistently outperforms the best methods while achieving superior inference efficiency.

2 Related Work

2.1 Multimodal Fake News Detection

Existing multimodal fake news detection methods can be broadly grouped into static multimodal fusion, robustness-oriented modeling, and evidence-aware grounding frameworks.

Early unimodal methods [18, 22] failed to detect sophisticated inter-modal contradictions, motivating multimodal detection. Current frameworks [23, 42, 45] usually fuse textual and visual representations through attention, contrastive learning, or mixture-of-experts modules. However, they often rely on static fusion over global embeddings, interpreting images superficially and applying the same visual processing depth to all samples.

These approaches face two limitations: (1) Weak visual grounding: Overlooking fine-grained details causes semantic gaps between textual claims and visual evidence [14, 20]. We address this by activating visual forensics only upon inter-modal discrepancies. (2) Noise injection: Indiscriminate fusion dilutes accuracy with irrelevant information [2, 19]. We mitigate this via a dynamic strategy regulating reasoning depth, selectively invoking visual analysis to avoid exhaustive processing.

2.2 Application of LLM in Fake News Detection

Multimodal LLMs enhance fake news detection [36, 29] by leveraging generative reasoning for active evidence extraction [5, 34]. Acting as forensic analysts, these models decompose complex narratives into atomic factual claims [9, 35] and generate interpretable visual descriptors, such as OCR, to bridge inter-modal semantic gaps.

Furthermore, LLMs facilitate consistency checking via reflection, explicitly reasoning about contra-

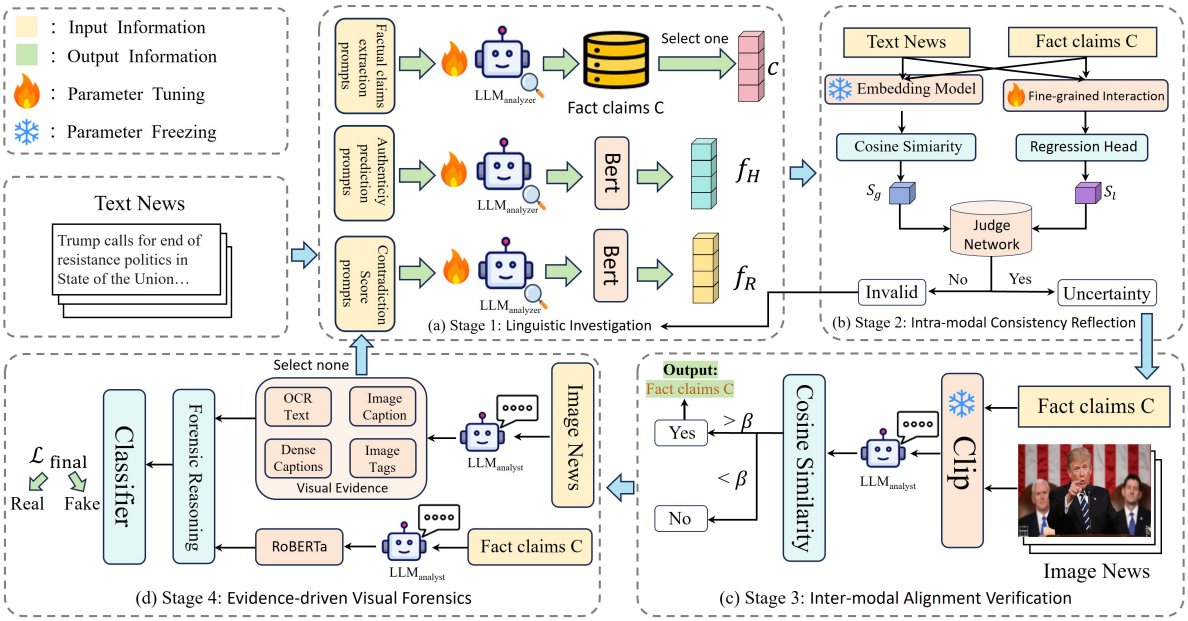

Figure 2: Overview of the DIVER framework featuring two explicit feedback loops. The pipeline initiates with linguistic investigation to verify intra-modal consistency. Subsequently, an inter-modal alignment gate dynamically regulates reasoning depth: samples exhibiting high semantic consensus are resolved immediately, while low-alignment cases trigger evidence-driven visual forensics, selectively activating visual analysis capabilities (e.g., OCR) to aggregate evidence for claim refinement.

dictions between claims and multimodal evidence [4, 38]. While susceptible to fabricating correlations without rigorous fine-grained visual grounding, integrating LLMs as dynamic engines within multi-stage frameworks [2, 4] remains a robust strategy for mitigating hallucinations.

3 Method

Unlike prior MFND methods that uniformly fuse visual features or use LLMs as single-pass reasoning modules, DIVER introduces a claim-level dynamic evidence routing paradigm. Its core novelty lies in explicitly separating evidence necessity estimation, evidence extraction, and evidence fusion: the model first verifies textual claims, then activates Evidence-driven Visual Forensics only when Inter-modal Alignment Verification indicates insufficient grounding, and finally integrates available branches through Uncertainty-Aware Dynamic Fusion.

Figure 2 shows the four-stage pipeline. DIVER extracts atomic claims from the news text, verifies their intra-modal consistency, estimates claim-image alignment, and selectively triggers evidence-driven visual forensics for low-alignment cases.

3.1 Linguistic Investigation

Given the raw news text T, we use a fine-tuned $\text{LLM}_{\text{analyst}}$ to extract factual claims and linguistic cues. Specifically, the LLM decomposes T into atomic factual claims $C = \{c_1, c_2, \ldots, c_k\}$ for subsequent inter-modal alignment, and analyzes authenticity-related and contradiction-related linguistic patterns. The generated analyses are encoded by BERT into dense vectors f_H and f_R:

$$C = \text{LLM}_{\text{analyst}}(P_{\text{extract}}, T), \quad (1)$$

$$f_H = \text{BERT}(\text{LLM}_{\text{analyst}}(P_{\text{auth}}, T)), \quad (2)$$

$$f_R = \text{BERT}(\text{LLM}_{\text{analyst}}(P_{\text{contra}}, T)). \quad (3)$$

3.2 Intra-modal Consistency Reflection

To reduce error propagation from hallucinated or shifted claims, we verify whether C faithfully preserves the semantics of T before introducing visual evidence.

Dual-branch verification. We measure the consistency between T and C from global and local perspectives. For global consistency, SFR-Embedding-Mistral [24] encodes T and the concatenated claims C, and their cosine similarity is used as S_g. For local consistency, a lightweight cross-attention module lets claim tokens attend to text tokens; the attended representation is mean-pooled and passed through a two-layer MLP to obtain S_l, capturing fine-grained omissions, perturbations, and contradictions. The final judge score

is:

$$p_{\text{judge}} = \sigma\left(\text{MLP}\left([S_g; S_l; |S_g - S_l|]\right)\right), \quad (4)$$

where p_{judge} denotes the probability that the extracted claims are faithful to the original text. The cross-attention module and judge network are trained on synthetic faithful/unfaithful claim pairs from the training split, with pretrained encoders frozen. Negative samples are constructed by entity replacement, number/date perturbation, omission, and contradiction injection.

Self-correction mechanism. If p_{judge} is below the validity threshold, DIVER converts the inconsistency into a diagnostic message and sends it back to the LLM for claim regeneration:

$$C^{(t+1)} = \text{LLM}_{\text{analyst}}(P_{\text{correct}}, T, C^{(t)}, \text{ErrorMsg}). \quad (5)$$

Here, ErrorMsg describes the inconsistency type, such as missing entities, semantic deviation, or factual contradiction. If the claims remain invalid after at most τ correction rounds, C is replaced with a conservative summary of T.

3.3 Inter-modal Alignment Verification

For verified claims $C = \{c_i\}_{i=1}^{k}$, we use a frozen CLIP model to estimate whether fine-grained visual evidence is necessary. This stage acts as an evidence-necessity estimator rather than a direct fake-news classifier. Instead of encoding all claims as one global sentence, we compute claim-wise image-text alignment scores:

$$s_i = \textbf{CLIP}_{\text{text}}(c_i) \cdot \textbf{CLIP}_{\text{img}}(I)^{\top}, \quad i = 1, \ldots, k. \quad (6)$$

However, not all textual claims are visually verifiable from a single image. Temporal, causal, source-attribution, or background claims may receive low CLIP scores even without visual contradiction. Therefore, we introduce a visual-verifiability mask $m_i^v \in \{0, 1\}$ to retain claims involving visible entities, objects, scenes, actions, or text:

$$\mathcal{V} = \{i \mid m_i^v = 1\}. \quad (7)$$

We then compute the inter-modal alignment score through conservative lower-bound aggregation over visually verifiable claims:

$$S_{\text{inter}} = \begin{cases} \min\limits_{i \in \mathcal{V}} s_i, & |\mathcal{V}| > 0, \\ +\infty, & |\mathcal{V}| = 0. \end{cases} \quad (8)$$

A validation-selected threshold β determines whether to activate visual forensics:

$$\text{Action} = \begin{cases} \text{Proceed to Stage 4}, & S_{\text{inter}} < \beta, \\ \text{Skip Forensics}, & S_{\text{inter}} \geq \beta. \end{cases} \quad (9)$$

Here, $S_{\text{inter}} < \beta$ indicates insufficient visual grounding for at least one visually verifiable claim, but it is not directly treated as evidence of fakeness. Instead, it only triggers Stage 4 to collect fine-grained visual evidence. Conversely, high-alignment samples bypass expensive visual forensics and proceed directly to uncertainty-aware fusion.

3.4 Evidence-driven Visual Forensics

When the inter-modal gate indicates insufficient visual grounding, DIVER activates an Evidence-driven Visual Forensics stage to resolve fine-grained image-text inconsistencies. Instead of exhaustively processing all visual signals, this stage performs claim-conditioned visual probing and extracts task-relevant evidence E_v for subsequent verification and refinement.

Adaptive Visual Evidence Probing. Given the verified claims $C = \{c_i\}_{i=1}^{k}$, we route each claim to the most relevant visual extractor according to its semantic type. Claims involving text, names, numbers, dates, or locations are routed to OCR; claims describing actions, scenes, or human-object interactions are routed to dense captioning; entity- or object-centric claims are routed to image tagging; otherwise, a general image captioning tool is used. This rule-based routing avoids invoking all tools for every sample while ensuring that each claim is checked by the most informative visual evidence source. The extracted OCR text, captions, dense descriptions, or tags are collected as visual evidence E_v.

Conditional Recursive Refinement. After obtaining visual evidence E_v, DIVER estimates whether the current claims are sufficiently supported by the extracted visual evidence. Instead of directly treating semantic distance as logical contradiction, we use it as a visual-text support

signal:

$$S_{\text{sup}} = \cos\left(\text{RoBERTa}(C), \text{RoBERTa}(E_v)\right). \tag{10}$$

If $S_{\text{sup}} < \gamma$, the visual evidence is regarded as insufficient to support the current claim set, and DIVER triggers an evidence-aware rollback. Importantly, this condition does not directly classify the sample as fake or contradictory; it only indicates that the current textual claims require refinement using grounded visual evidence. In this case, E_v is injected as grounded context to guide claim re-extraction:

$$C_{\text{refined}} = \text{LLM}_{\text{analyst}}(P_{\text{re-extract}}, T, E_v). \tag{11}$$

This refinement step revises ambiguous or visually unsupported claims by anchoring the LLM reasoning in concrete visual artifacts, thereby reducing error propagation.

Semantic Evidence Consolidation. After refinement, the heterogeneous visual findings are distilled by the LLM into a compact evidence summary and encoded as the visual evidence feature $\mathbf{f}_v$. If Stage 4 is skipped by the gate in Eq. 9, we set $\mathbf{f}_v = \mathbf{0}$ and assign the visual mask $m_v = 0$; otherwise, $m_v = 1$. This design ensures that visual evidence contributes to the final fusion only when it is explicitly activated.

3.5 Uncertainty-Aware Dynamic Fusion

Our Uncertainty-Aware Dynamic Fusion (UDF) aggregates features $F = \{f_t, f_H, f_R, f_v\}$ (where f_t denotes the text embedding of verified claims C). UDF robustly handles intermittent missing modalities from the skipping mechanism. Crucially, when visual forensics is bypassed (e.g., for generic images), UDF adaptively shifts reliance to textual branches via uncertainty modeling.

First, we derive a refined feature $\hat{f}_k$ for each branch, weighted by its estimated reliability μ_k. To make the contribution of activated visual evidence explicitly controllable, we introduce a visual-evidence scaling coefficient λ before masked attention:

$$\begin{aligned} \mu_k &= \sigma\big(\mathbf{w}_\mu^T \tanh(\mathbf{W}_p f_k + \mathbf{b}_p)\big), \\ \hat{f}_k &= \begin{cases} \lambda \mu_k f_k, & k = v, \\ \mu_k f_k, & k \neq v. \end{cases} \end{aligned} \tag{12}$$

Here, $k = v$ denotes the visual-evidence branch. The coefficient λ controls the pre-attention scale of activated visual evidence, while the final branch contribution is still adaptively determined by the subsequent masked attention. We select λ on the validation split and set $\lambda = 0.40$ in all experiments.

Conventional fusion misinterprets zero-padded features from bypassed stages as signals. To resolve this, UDF employs a masked attention mechanism guided by a binary state $m_k \in \{0, 1\}$ for each modality k. A learnable query $\mathbf{q}_{task}$ guides the attention:

$$\begin{aligned} e_k &= \frac{(\mathbf{W}_Q \mathbf{q}_{task})^T (\mathbf{W}_K \hat{f}_k)}{\sqrt{d_h}}, \\ \alpha_k &= \text{Softmax}(e_k + \mathcal{M}_k). \end{aligned} \tag{13}$$

where the mask $\mathcal{M}_k \in \{0, -\infty\}$ is defined as $-\infty$ if $m_k = 0$ (skipped) and 0 otherwise. This strict masking ensures missing modalities (e.g., $f_v = \mathbf{0}$) exert zero influence, maintaining mathematical rigor.

The aggregated representation $f_{\text{final}} = \sum_k \alpha_k (\mathbf{W}_V \hat{f}_k)$ is fed into an MLP to predict news veracity $\hat{y} = \sigma(\mathbf{W}_c f_{\text{final}} + b_c)$. We optimize the framework using binary cross-entropy loss:

$$\mathcal{L}_{\text{final}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{14}$$

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate DIVER on three benchmarks: Weibo [33], Weibo21 [25], and GossipCop [23]. Weibo contains 7,532 training and 1,996 test articles; Weibo21 contains 9,127 articles in total; and GossipCop includes 10,010 training and 2,830 test instances. The class distributions are 3,749/3,783 real/fake for Weibo training, 996/1,000 for Weibo testing, 4,640/4,487 for Weibo21, and 7,974/2,036 training plus 2,285/545 testing real/fake instances for GossipCop.

Baselines. We compare DIVER with four categories of methods: (1) multimodal multi-domain methods, including FND-CLIP [43], MIMoE-FND [23], KEN [45], MViR [21], and DAUD [38]; (2) multimodal single-domain methods, including RaCMC [41], CAFE [32], BMR [40], and SEER [44]; (3) LLM-based reasoning methods, including GLPN-LLM [17], INSIDE [34], LIFE [31], and FactGuard [14]; and (4) general LVLM/LLM backbones, including LLaVA-1.5-7B, InternVL2-8B, Qwen3-VL-7B, and Llama3.1-8B. For fair comparison, we use the strongest official configurations

Table 1: Comparison of DIVER with the latest multi-domain fake news detection methods. The best and second-best results are highlighted in red and blue, respectively. The superscripts in the DIVER row indicate the performance improvement over the best baseline. All results are reported as percentages (%).

		Weibo				Weibo21				GossipCop			
Category	**Method**	**Acc**	**F1-Fake**	**F1-Real**	**AUC**	**Acc**	**F1-Fake**	**F1-Real**	**AUC**	**Acc**	**F1-Fake**	**F1-Real**	**AUC**
Multimodal multi-domain	FND-CLIP (ICME'23)	90.7	90.8	90.7	95.3	94.3	94.0	94.6	96.2	88.0	63.8	92.8	87.1
	MIMoE-FND (WWW'25)	92.8	92.8	92.8	97.2	95.6	95.5	95.7	97.7	89.5	69.8	93.6	87.9
	KEN (ACM MM'25)	93.5	93.5	93.4	96.7	93.5	93.7	93.2	97.1	88.1	64.6	92.8	87.3
	MViR (ICASSP'26)	92.4	92.0	92.8	95.6	95.2	95.0	94.8	96.8	89.5	69.2	93.7	87.5
	DAUD (WWW'26)	92.7	92.9	93.5	96.3	94.6	95.2	95.8	97.4	89.2	69.5	92.0	86.2
Multimodal single-domain	RaCMC (AAAI'25)	91.5	91.7	91.4	92.1	94.2	93.8	94.3	96.2	87.9	64.1	92.7	83.8
	CAFE (WWW'22)	84.0	84.2	83.7	89.2	88.2	88.5	87.6	90.9	86.7	58.7	92.1	85.2
	BMR (AAAI'23)	91.8	91.4	90.4	95.4	92.9	92.7	92.5	96.2	89.5	69.1	87.6	88.1
	SEER (SMC'25)	92.9	92.8	93.9	93.4	93.2	92.7	92.5	96.0	89.3	67.3	87.1	87.5
Reasoning LLMs	INSIDE (CIKM'25)	88.1	68.4	93.2	91.0	89.6	81.6	83.2	87.1	90.0	70.7	93.4	88.1
	GLPN-LLM (ACL'25)	92.0	93.9	93.9	95.4	92.5	93.7	92.4	95.9	89.0	68.2	93.3	86.4
	LIFE (WWW'26)	92.4	94.0	92.1	95.8	92.8	93.5	93.4	97.5	86.4	70.1	86.6	85.5
	FactGuard (AAAI'26)	89.3	90.6	88.7	92.6	80.4	77.7	82.4	93.8	89.2	67.5	93.5	85.8
Large Language Models	LLaVA-1.5-7B	93.1	93.3	93.0	97.0	93.8	93.9	93.7	96.8	88.5	66.5	93.0	86.8
	InternVL2-8B	93.6	93.8	93.4	96.9	94.2	95.5	94.3	97.0	89.0	68.0	93.2	87.0
	Qwen3-VL-7B	89.5	89.0	89.8	94.5	90.5	90.2	90.8	95.0	89.8	70.5	93.6	88.0
	Llama3.1-8B	87.2	86.5	87.8	92.5	88.4	88.0	88.7	93.2	88.2	65.4	92.5	86.5
Ours	**DIVER**	**95.8**	**95.2**	**95.7**	**98.2**	**97.7**	**96.6**	**97.0**	**98.8**	**91.6**	**72.3**	**95.4**	**89.3**
	Improv.	2.46%↑	1.28%↑	1.92%↑	1.03%↑	2.20%↑	1.15%↑	1.36%↑	1.13%↑	1.78%↑	2.26%↑	1.71%↑	1.36%↑

whenever available, including reported results, fine-tuned settings, prompt designs, and comparable 7B/8B-scale backbones. Re-implemented baselines follow their released protocols rather than weakened zero-shot settings. Additional backbone-, evidence-, budget-, and open-source-matched controls are provided in Appendix E.

Metrics. We report accuracy, F1-score, and AUC. Following common practice, accuracy is the principal metric for GossipCop, while F1-Real is emphasized for Weibo and Weibo21 to reflect real-news detection performance. Higher values indicate better performance.

Implementation Details. We extract visual features using CLIP with a ViT-B/32 backbone [28, 7], and encode text with `bert-base-chinese` [6] for Weibo/Weibo21 and `bert-base-uncased` [6] for GossipCop, with a maximum length of 197 tokens. We fine-tune Llama3-8B as $\text{LLM}_{\text{analyst}}$ on official training splits [37, 30, 12], converting each news instance into instruction-style supervision for claim extraction, consistency verification, and evidence-aware reasoning. No test labels or test-set rationales are used, and evaluation uses only raw news text and images. During inference, the LLM temperature is set to 0.4, with a context length of 4096 tokens, extended to 8192 when visual evidence is injected. We set the maximum self-correction rounds to $\tau = 2$, the contradiction threshold to $\gamma = 0.45$, and the validation-selected inter-modal gating threshold to $\beta = 0.29$. For Stage 3.4, we use four Anthropic API-based image-to-text tools, including OCR, Image Captioning, Dense Captioning, and Image Tagging, with the same model version and prompt template across datasets. The intra-modal judge is trained on the training split before the full DIVER pipeline and kept fixed afterward. The framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU for 50 epochs with early stopping.

4.2 Overall Performance

We evaluate our framework against 16 strong baselines on three widely used datasets: Weibo, Weibo21 and GossipCop. The comprehensive experimental results for all methods are presented in Table 1.

From the results, we observe that our DIVER framework consistently outperforms existing best methods on all datasets, demonstrating its effectiveness. Specifically, DIVER achieves accuracies of 95.8%, 97.7%, and 91.6% on Weibo, Weibo21, and GossipCop, respectively, which are 2.46%, 2.20%, and 1.78% higher than the previous best methods (i.e., KEN and MIMoE-FND).

We further analyze the results from three perspectives:

Comparison with Multimodal Baselines: DIVER achieves the best performance among the compared baselines, reaching 95.8%, 97.7%, and

Table 2: Ablation study results showing performance drops for different model variants. We report accuracy, F1-Fake, F1-Real and AUC. Module Abbreviations: Embed. (Embedding Model), Vis.Evid. (Visual Evidence), Feedbk (Feedback Mechanism), Facts (Factual Claims Extraction).

Model	Weibo				Weibo21				GossipCop			
	Acc	**F1-Fake**	**F1-Real**	**AUC**	**Acc**	**F1-Fake**	**F1-Real**	**AUC**	**Acc**	**F1-Fake**	**F1-Real**	**AUC**
DIVER (Ours)	95.8	95.2	95.7	98.2	97.7	96.6	97.0	98.8	91.6	72.3	95.4	89.3
Linguistic Investigation												
- w/o **Prompts**	92.4 ↓3.4%	92.3 ↓2.9%	92.6 ↓3.1%	94.2 ↓4.0%	93.5 ↓4.2%	93.2 ↓3.4%	93.6 ↓3.4%	93.1 ↓5.7%	88.1 ↓3.5%	68.7 ↓3.6%	91.2 ↓4.2%	85.8 ↓3.5%
- w/o $\text{LLM}_{\text{analyst}}$	92.5 ↓3.3%	92.2 ↓3.0%	92.9 ↓2.8%	93.8 ↓4.4%	93.1 ↓4.6%	93.4 ↓3.2%	94.0 ↓3.0%	92.6 ↓6.2%	87.8 ↓3.8%	67.9 ↓4.4%	92.4 ↓3.0%	85.3 ↓4.0%
Intra-modal Consistency Reflection												
- w/o **Embed.**	92.9 ↓2.9%	93.1 ↓2.1%	92.6 ↓3.1%	94.6 ↓3.6%	94.5 ↓3.2%	93.8 ↓2.8%	94.1 ↓2.9%	93.3 ↓5.5%	88.4 ↓3.2%	69.1 ↓3.2%	90.8 ↓4.6%	85.9 ↓3.4%
- w/o **Judge**	93.0 ↓2.8%	92.5 ↓2.7%	92.4 ↓3.3%	94.2 ↓4.0%	94.3 ↓3.4%	93.5 ↓3.1%	93.3 ↓3.7%	92.8 ↓6.0%	88.0 ↓3.6%	67.7 ↓4.6%	92.2 ↓3.2%	85.4 ↓3.9%
Inter-modal Alignment Verification												
- w/o $\text{LLM}_{\text{analyst}}$	91.2 ↓4.6%	91.9 ↓3.3%	92.2 ↓3.5%	93.5 ↓4.7%	91.0 ↓6.7%	92.5 ↓4.1%	92.1 ↓4.9%	92.7 ↓6.1%	86.8 ↓4.8%	68.3 ↓4.0%	90.1 ↓5.3%	85.2 ↓4.1%
- w/o **CLIP**	90.4 ↓5.4%	91.6 ↓3.6%	91.7 ↓4.0%	94.4 ↓3.8%	91.7 ↓6.0%	92.5 ↓4.1%	93.4 ↓3.6%	93.3 ↓5.5%	87.5 ↓4.1%	69.2 ↓3.1%	90.9 ↓4.5%	86.2 ↓3.1%
Evidence-driven Visual Forensics												
- w/o **Facts**	90.8 ↓5.0%	91.7 ↓3.5%	92.4 ↓3.3%	93.6 ↓4.6%	92.1 ↓5.6%	92.6 ↓4.0%	93.3 ↓3.7%	92.5 ↓6.3%	87.4 ↓4.2%	68.8 ↓3.5%	90.5 ↓4.9%	85.1 ↓4.2%
- w/o **Vis.Evid.**	91.6 ↓4.2%	91.4 ↓3.8%	92.2 ↓3.5%	93.1 ↓5.1%	93.1 ↓4.6%	92.3 ↓4.3%	92.9 ↓4.1%	92.2 ↓6.6%	87.2 ↓4.4%	68.6 ↓3.7%	90.8 ↓4.6%	85.0 ↓4.3%
- w/o **Feedbk**	90.7 ↓5.1%	90.9 ↓4.3%	91.8 ↓3.9%	93.7 ↓4.5%	92.6 ↓5.1%	92.0 ↓4.6%	92.5 ↓4.5%	92.4 ↓6.4%	86.9 ↓4.7%	69.3 ↓3.0%	91.3 ↓4.1%	85.6 ↓3.7%

91.6% accuracy on Weibo, Weibo21, and GossipCop, respectively. This consistent superiority demonstrates that our dynamic gating mechanism effectively mitigates multimodal noise by selectively filtering irrelevant background details, addressing the accuracy bottlenecks caused by indiscriminate fusion in baselines like RaCMC.

Comparison with LLM-reasoning Methods: Unlike standard LLM-based approaches (e.g., GLPN-LLM, LIFE) that often struggle with class imbalance, DIVER delivers substantial gains in both F1-Fake (+2.26% on GossipCop) and F1-Real (+1.71% on GossipCop). This balanced performance indicates that our iterative visual feedback strategy provides more refined reasoning than single-pass extraction, effectively reducing both false positives and false negatives.

Robustness of Reflections: The robustness of our framework is evidenced by its high AUC scores across all datasets, peaking at 98.2% on Weibo and 98.8% on Weibo21. These results highlight the efficacy of our collaborative reflections in ensuring semantic consistency, which minimizes high-confidence hallucinations and enhances the model's capability to reliably distinguish between authentic and fabricated news.

4.3 Ablation Analysis

To evaluate the contribution of each component in DIVER, we conduct ablation studies on Weibo, Weibo21, and GossipCop, as shown in Table 2. For each variant, the removed module is replaced with the simplest compatible alternative to keep the pipeline executable. Specifically, w/o Prompts uses generic task instructions; w/o $\text{LLM}_{\text{analyst}}$ replaces instruction-tuned analysis with direct text encoding; w/o Judge skips intra-modal faithfulness verification; w/o CLIP disables the inter-modal alignment gate and forwards all samples to visual reasoning; w/o Facts uses the original text instead of extracted atomic claims; w/o Vis.Evid. removes OCR/caption/tag evidence; and w/o Feedbk disables recursive evidence-aware refinement.

Effect of linguistic investigation. Removing the designed prompts or $\text{LLM}_{\text{analyst}}$ consistently degrades performance, with accuracy drops of about 2.5%–3.8% across datasets. This shows that instruction-guided claim extraction and linguistic reasoning provide a strong textual foundation before multimodal verification.

Effect of intra-modal reflection. Removing the embedding-based consistency branch or the judge network also causes clear performance drops. This confirms that verifying the faithfulness of extracted claims is necessary for reducing hallucinated or semantically shifted claims before introducing visual evidence.

Effect of inter-modal alignment and visual forensics. The largest degradation appears when removing CLIP-based alignment, factual claims, visual evidence, or feedback refinement. In particular, disabling the CLIP gate weakens the model's ability to distinguish visually consistent samples from ambiguous ones, causing either unnecessary visual noise or insufficient grounding. Similarly, removing visual evidence or feedback prevents DIVER

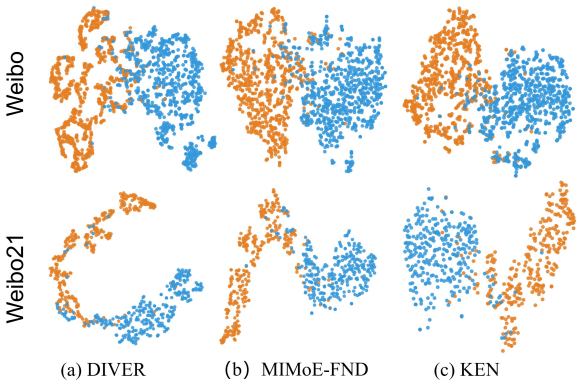

Figure 3: T-SNE visualization of test set features. Dots with the same color indicate the same label.

from resolving fine-grained image-text contradictions. These results validate the necessity of dynamic gating, claim-level evidence extraction, and recursive refinement in DIVER.

4.4 Visualization Results

Figure 3 shows T-SNE visualizations of features from DIVER, MIMoE-FND [23], and KEN [45] on the Weibo and Weibo21 test sets. Compared to the baselines, DIVER produces fewer fake news outliers and less overlap between real and fake news embeddings, confirming its superior performance.

On Weibo21, DIVER's features form multiple, separated subclusters, unlike the single clusters on Weibo. This suggests Weibo21 has varying topics and that DIVER not only distinguishes authenticity but also captures deep, event-level semantic information, spatially distinguishing events.

4.5 Inference Efficiency and Cost Analysis

To evaluate practical applicability, we compare DIVER with recent LLM-based methods (GLPN-LLM, INSIDE, and LIFE) in terms of inference latency (Table 3) and amortized computational overhead (Table 4).

Table 3: Average response time (s) for verifying a single news sample using different LLM-based methods. DIVER achieves the fastest inference speed.

Model	Weibo	Weibo21	GossipCop	Avg.
GLPN-LLM	5.35	5.95	5.88	5.73
INSIDE	4.53	5.12	5.25	4.97
LIFE	6.21	7.45	7.60	7.09
DIVER (Ours)	**3.55**	**4.20**	**4.32**	**4.12**

Inference Latency. Table 3 shows that DIVER achieves the fastest inference, reducing average reasoning time by 2.97 s over LIFE and 0.85 s over INSIDE. This speedup comes from Inter-modal Alignment Verification, which allows straightforward samples to bypass expensive visual forensics while preserving deeper reasoning for ambiguous cases.

Table 4: Comparison of computational overhead and performance. DIVER reduces amortized model/tool invocations and token costs via dynamic gating while achieving the best accuracy.

Method	Acc (%)			Efficiency	
	Weibo	Weibo21	GossipCop	Invoc.	Tok. (k)
INSIDE	88.1	89.6	90.0	2.8	2.3
GLPN-LLM	92.0	92.5	89.0	2.6	1.9
LIFE	92.4	92.8	86.4	3.2	3.2
DIVER	**95.8**	**97.7**	**91.6**	**2.1**	**1.5**

Token and Invocation Efficiency. Table 4 reports the cost-benefit trade-off. DIVER uses only 1.5k tokens per sample, 21% fewer than GLPN-LLM at 1.9k, and requires only 2.1 amortized model/tool invocations, fewer than GLPN-LLM (2.6) and LIFE (3.2). The invocation number is averaged over all test samples under the dynamic routing protocol: skipped visual tools contribute zero invocations, while additional visual evidence extraction is activated only for low-alignment samples. These gains do not compromise performance, as DIVER achieves the best accuracy across all datasets.

5 Conclusion

This paper presents DIVER, a dynamic evidence-grounded reasoning framework for multimodal fake news detection. Rather than treating visual information as a uniformly fused feature, DIVER shows that visual evidence should be activated selectively according to the grounding status of textual claims. By combining faithful claim construction, inter-modal discrepancy detection, evidence-driven visual forensics, and uncertainty-aware fusion, DIVER provides a more adaptive alternative to static multimodal fusion and single-pass LLM reasoning. Experiments on Weibo, Weibo21, and GossipCop demonstrate that this dynamic reasoning paradigm improves both detection accuracy and inference efficiency. These results suggest that effective multimodal misinformation detection requires not only stronger representations, but also better control over when and how visual evidence is introduced.

6 Limitations.

We would like to further explore the following limitations: ① enhancing the framework's generalizability to more diverse social media domains, ② evaluating its scalability in more complex real-world scenarios and its robustness against unseen or adaptive adversarial attacks, and ③ investigating lightweight model distillation techniques to better adapt the framework to resource-constrained edge devices. ④ The datasets may contain news images, public figures, social-media text, or potentially misleading and sensitive content. We use these data only for research purposes, do not attempt to identify private individuals, and report only aggregate experimental results.

References

[1] Jawaher Alghamdi, Suhuai Luo, and Yuqing Lin. 2024. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83(17):51009–51067.

[2] Y. Bu, Q. Sheng, J. Cao, S. Wang, P. Qi, Y. Shi, and B. Hu. 2025. Enhancing fake news video detection via llm-driven creative process simulation. *arXiv preprint arXiv:2510.04024*.

[3] R. Chakraborty, M. Bhavsar, S.K. Dandapat, and J. Chandra. 2022. Detecting stance in tweets: A signed network based approach. *arXiv preprint arXiv:2201.07472*.

[4] R. Chalehchaleh, R. Farahbakhsh, and N. Crespi. 2025. Addressing data scarcity in multilingual fake news detection: an llm-based dataset augmentation approach. *Social Network Analysis and Mining*, 15(1):1–16.

[5] B. Chen, Z. Yue, S. Chen, Z. Wang, Y. Liu, P. Li, and Y. Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint arXiv:2503.10200*.

[6] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[8] Don Fallis and Kay Mathiesen. 2025. Fake news is counterfeit news. *Inquiry*, 68(10):3191–3210.

[9] F. Farhangian, L. A. Ensina, G. D. C. Cavalcanti, and R. M. O. Cruz. 2025. Dres: Fake news detection by dynamic representation and ensemble selection. *arXiv preprint arXiv:2509.16893*.

[10] Faramarz Farhangian, Rafael MO Cruz, and George DC Cavalcanti. 2024. Fake news detection: Taxonomy and comparative study. *Information Fusion*, 103:102140.

[11] Aaron M French, Veda C Storey, and Linda Wallace. 2025. The impact of cognitive biases on the believability of fake news. *European Journal of Information Systems*, 34(1):72–93.

[12] 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*.

[13] 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.

[14] Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2025. Factguard: Event-centric and commonsense-guided fake news detection. *arXiv preprint arXiv:2511.10281*.

[15] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 22105–22113.

[16] Bo Hu, Zhendong Mao, and Yongdong Zhang. 2025. An overview of fake news detection: From a new perspective. *Fundamental research*, 5(1):332–346.

[17] S. Hu, J. Hu, and H. Zhang. 2025. Synergizing llms with global label propagation for multimodal fake news detection. *arXiv preprint arXiv:2506.00488*.

[18] Y. Jiang, T. Wang, X. Xu, Y. Wang, X. Song, and D. Maynard. 2025. Cross-modal augmentation for few-shot multimodal fake news detection. *Engineering Applications of Artificial Intelligence*, 142:109931.

[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. 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.

[20] Jian Lang, Rongpei Hong, Ting Zhong, Yong Wang, and Fan Zhou. 2026. Nip rumors in the bud: Retrieval-guided topic-level adaptation for test-time fake news video detection. *arXiv preprint arXiv:2601.11981*.

[21] Haochen Liang, Xinqi Su, Jun Wang, Chaomeng Chen, and Zitong Yu. 2026. Mvir: Multi-view visual-semantic representation for fake news detection. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12192–12196.

[22] X. Liu, P. Li, H. Huang, Z. Li, X. Cui, J. Liang, L. Qin, W. Deng, and Z. He. 2024. Fakenewsgpt4: advancing multimodal fake news detection through knowledge-augmented lvlms. *CoRR*.

[23] Y. Liu, Y. Liu, Z. Li, R. Yao, Y. Zhang, and D. Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.

[24] R. Meng, Y. Liu, S.R. Joty, C. Xiong, Y. Zhou, and S. Yavuz. 2024. Sfrembedding-mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog.

[25] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.

[26] Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

[27] Jan Pfänder and Sacha Altay. 2025. Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. *Nature human behaviour*, 9(4):688–699.

[28] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[29] C. Tian, Q. Ho, and X. Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. *arXiv preprint arXiv:2508.19633*.

[30] Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

[31] C. Wang, M. Gao, Z. Wang, J. Yin, K. Shu, and C. Lin. 2025. Prompt-induced linguistic fingerprints for llm-generated fake news detection. *arXiv preprint arXiv:2508.12632*.

[32] 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.

[33] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857.

[34] Z. Wang, Q. Sheng, D. Wang, B. Hu, and J. Cao. 2025. Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection. *arXiv preprint arXiv:2509.01660*.

[35] Shicai Wei, Chunbo Luo, and Yang Luo. 2025. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22879–22888.

[36] Y. Wei, H. Zhou, S. Yuan, M. Chen, H. Shi, Z. Jia, L. Wang, and X. He. 2025. Deepmsd: Advancing multimodal sarcasm detection through knowledge-augmented graph reasoning. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6413–6423.

[37] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

[38] Xuankai Yang, Yan Wang, Jiajie Zhu, Pengfei Ding, Hongyang Liu, Xiuzhen Zhang, and Huan Liu. 2026. Cross-domain fake news detection on unseen domains via llm-based domain-aware user modeling. *arXiv preprint arXiv:2602.01726*.

[39] Jingyuan Yi, Zeqiu Xu, Tianyi Huang, and Peiyang Yu. 2025. Challenges and innovations in llm-powered fake news detection: A synthesis of approaches and future directions. In *Proceedings of the 2025 2nd international conference on generative artificial intelligence and information security*, pages 87–93.

[40] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5384–5392.

[41] X. Yu, Z. Sheng, W. Lu, X. Luo, and J. Zhou. 2025. Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 986–994.

[42] H. Zhou, Y. Wei, J. Yang, and Z. Zhang. 2025. Towards robust and reliable multimodal fake news detection with incomplete modality. *arXiv preprint arXiv:2510.05839*.

[43] Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE.

[44] P. Zhu, Y. Jing, L. Cheng, B. Chen, X. Cui, L. Wu, and K. Tang. 2025. Seer: Semantic enhancement and emotional reasoning network for multimodal fake news detection. *arXiv preprint arXiv:2507.13415*.

[45] P. Zhu, Y. Jing, L. Cheng, K. Tang, and Y. Guo. 2025. Ken: Knowledge augmentation and emotion guidance network for multimodal fake news detection. *arXiv preprint arXiv:2507.09647*.

A Additional Case Study: Model Explainability

To further demonstrate the explainability of DIVER, we provide representative case studies showing how the model adaptively decides whether deeper visual evidence is necessary. As shown in Table 5, DIVER avoids unnecessary evidence extraction for visually consistent samples, while activating more fine-grained forensic tools when cross-modal inconsistency is detected.

Specifically, Cases 1 and 2 correspond to real news samples where the image content is broadly consistent with the textual claim. In these cases, DIVER can rely on lightweight visual-textual verification and bypass part of the deeper evidence-driven modules, such as OCR or dense forensic extraction. This behavior reflects the cognitive economy principle of the framework: the model does not invoke expensive tools when shallow cross-modal evidence is already sufficient.

In contrast, Case 3 presents an adversarial fake-news example, where the textual claim states that the image depicts Brad Pitt attending the Golden Globe Awards. However, DIVER triggers the evidence-driven module and extracts fine-grained visual clues, including the OCR result "TEAM USA" and tags such as "Sports", "Olympics", and "Team USA". These visual cues directly contradict the textual claim and provide interpretable evidence for the final fake-news judgment. This case shows that DIVER does not merely rely on global image-text similarity, but performs targeted evidence extraction when potential inconsistency is detected.

Scope of the cross-dataset evaluation. The above transfer setting mainly evaluates cross-dataset robustness within the Chinese multimodal fake-news detection scenario. Although Weibo and Weibo21 differ in event distributions, topic patterns, and dataset-specific biases, they share the same language environment. Therefore, these results should be interpreted as evidence of Chinese cross-dataset generalization rather than a complete evaluation of cross-lingual transfer.

Cross-lingual transfer as a stricter stress test. A stricter setting is Chinese–English transfer, e.g., Weibo/Weibo21 $\rightarrow$ GossipCop and GossipCop $\rightarrow$ Weibo/Weibo21. This setting is substantially more challenging because language shift, platform shift, topic shift, and class-prior shift occur simultaneously. In particular, GossipCop focuses on English entertainment news and is more imbalanced than Weibo and Weibo21, making direct transfer less comparable to the Chinese cross-dataset setting. We therefore regard cross-lingual transfer as a complementary stress test rather than a replacement for the standard source-only cross-dataset evaluation.

B Notation

Table 6 summarizes the main symbols used in DIVER.

C Reproducibility and Fairness Details of Visual Evidence Tools

To improve reproducibility and fairness, we provide additional details of the visual evidence tools used in Evidence-driven Visual Forensics. The four tool names, i.e., OCR, Image Captioning, Dense Captioning, and Image Tagging, refer to four functional prompting roles implemented with the same image-to-text API model version. Across all datasets, we keep the model version, decoding setting, prompt template, and output format fixed. No ground-truth label, dataset identity, test-set rationale, or dataset-specific label hint is provided to any visual evidence tool.

Backend-agnostic evidence interface. Although our implementation uses Anthropic API-based image-to-text models, DIVER is not tied to a specific commercial backend. The visual tools instantiate a unified evidence interface whose role is to produce observable, claim-conditioned visual evidence rather than veracity predictions. The backend receives only the raw image and the verified textual claim, and does not access the ground-truth label, dataset identity, final prediction target, or any label-related hint. Therefore, the external model is used as an evidence extractor rather than a fake-news classifier.

Prompt templates. For each activated tool role,

Table 5: Explainability case study. Consistent samples, i.e., Cases 1 and 2, bypass deep forensics marked as "Not Reached" for efficiency. In contrast, the adversarial Case 3 triggers the evidence-driven module, where specific visual clues, such as OCR result "TEAM USA", are extracted to explicitly refute the contradictory text claim.

	Keys	Case 1 (Real)	Case 2 (Real)	Case 3 (Fake)
News	Image News			
	Text News	Host Andy Cohen interviews actresses Jane Fonda and Lily Tomlin	Kate Middleton joins Meghan Markle for Christmas service	Brad Pitt attends the annual Golden Globe Awards
Method	OCR	Not Reached	Not Reached	TEAM USA on scarves
	Caption	Andy Cohen interviews guests on his show	Royal family members walk side by side	Not Reached
	Dense Captions	Not Reached	Two women wearing hats and coats walking	Smiling people wearing white sports jackets
	Tags	Talk show, Interview, TV Host	Not Reached	Sports, Olympics, Team USA

Table 6: Main notation used in DIVER.

Type	Symbol	Description
Input Claims	$x = (T, I)$	News text T and paired image I.
	$C = \{c_i\}_{i=1}^{k}$	Extracted factual claims; k is the claim number.
Features	$\mathbf{f}_t \in \mathbb{R}^d$	Verified claim feature.
	$\mathbf{f}_H \in \mathbb{R}^d$	Linguistic style/authenticity feature.
	$\mathbf{f}_R \in \mathbb{R}^d$	Linguistic contradiction feature.
	$\mathbf{f}_v \in \mathbb{R}^d$	Visual evidence feature.
	$\mathbf{F}$	Feature set $\{\mathbf{f}_t, \mathbf{f}_H, \mathbf{f}_R, \mathbf{f}_v\}$.
Scores	S_g	Global text-claim consistency score.
	S_l	Fine-grained local consistency score.
	S_{inter}	Image-text alignment score.
	S_{sup}	Visual-text support score for evidence-aware claim refinement.
Thresholds	β	Validation-selected threshold controlling whether Stage 4 is activated.
	γ	Threshold for evidence-aware rollback.
Fusion	$m_k \in \{0, 1\}$	Availability mask of branch k.
	$\mathbf{q}_{task}$	Learnable task query.
	$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$	Projection matrices in masked attention.
	e_k, α_k	Attention logit and normalized weight.
	$\mathcal{M}_k$	Attention mask, either 0 or $-\infty$.

we use a fixed prompt template. The input consists only of the raw image and the verified textual claim. The templates are summarized below:

OCR. Given the image and the claim, extract all visible textual evidence that may support or contradict the claim. Return only readable text spans and their brief visual context. If no readable text is visible, return None.

Image Captioning. Given the image and the claim, describe the overall visual content in one or two concise sentences. Focus on visually grounded facts and avoid inferring identities or events not visible in the image.

Dense Captioning. Given the image and the claim, describe fine-grained visual regions, actions, interactions, and scene details that may be relevant to the claim. Return short evidence phrases.

Image Tagging. Given the image and the claim, list salient objects, entities, attributes, scenes, and symbols that are visually observable and potentially relevant to the claim.

Fairness control. To avoid dataset-specific tool engineering, all datasets share the same image-to-text model version, decoding configuration, prompt templates, and output format. The four tool roles differ only in their functional prompts, while using the same backend and inference setting. The extracted evidence is normalized into the same textual evidence format and processed by the same refinement and fusion modules. This protocol ensures that the performance gain mainly comes from DIVER's claim-conditioned evidence routing and dynamic fusion, rather than from dataset-specific prompts or label-aware visual tools.

Post-processing and failure handling. The raw tool outputs are normalized into a unified textual evidence list E_v. Empty outputs are preserved as None rather than being manually completed. If an API call fails or returns an invalid format, we retry once with the same prompt and decoding setting.

Table 7: Routing protocol and output format of visual evidence tools.

Claim Type	Tool Role	Prompt Objective	Output Format
Text, names, numbers, dates, locations	OCR	Extract visible text relevant to the claim, including signs, subtitles, logos, timestamps, or written names.	A list of text spans with brief visual context; return None if no readable text exists.
Actions, scenes, interactions	Dense Captioning	Describe localized visual events, human-object interactions, and scene-level relations that can support or contradict the claim.	Short evidence phrases describing actions, relations, and visible scene cues.
Entities, objects, attributes	Image Tagging	Identify salient entities, objects, clothing, symbols, scenes, and attribute tags related to the claim.	A ranked list of concise tags with optional attribute descriptions.
General visual context	Image Captioning	Produce a compact image-level description when the claim does not match a specialized tool category.	One or two concise natural-language captions.

If the second attempt still fails, we fall back to the general image captioning role and mark the corresponding evidence source as unavailable in the evidence summary. This prevents failed tools from introducing unsupported visual evidence.

Reproducibility without commercial re-querying. To make the results reproducible even when commercial APIs are updated, we release the complete prompt templates, decoding settings, model-version information, and cached visual evidence outputs for all evaluated samples. As a result, researchers can reproduce the reported evidence refinement and final fusion results without re-querying the commercial API. The API is only required when regenerating visual evidence from raw images.

Open-source instantiation. DIVER can also be instantiated with open-source OCR, captioning, dense-captioning, or tagging models without changing the routing, refinement, or fusion architecture. This backend-agnostic design separates the methodological contribution of dynamic evidence control from the choice of a particular visual recognition engine, making DIVER applicable to both commercial and fully open-source deployment settings.

Prompt templates. For each activated tool role, we use a fixed prompt template. The input consists only of the raw image and the verified textual claim. The templates are summarized below:

> **OCR.** Given the image and the claim, extract all visible textual evidence that may support or contradict the claim. Return only readable text spans and their brief visual context. If no readable text is visible, return None.
>
> **Image Captioning.** Given the image and the claim, describe the overall visual content in one or two concise sentences. Focus on visually grounded facts and avoid inferring identities or events not visible in the image.
>
> **Dense Captioning.** Given the image and the claim, describe fine-grained visual regions, actions, interactions, and scene details that may be relevant to the claim. Return short evidence phrases.
>
> **Image Tagging.** Given the image and the claim, list salient objects, entities, attributes, scenes, and symbols that are visually observable and potentially relevant to the claim.

Post-processing and failure handling. The raw tool outputs are normalized into a unified textual evidence list E_v. Empty outputs are preserved as None rather than being manually completed. If the API call fails or returns an invalid format, we retry once with the same prompt and decoding setting. If the second attempt still fails, we fall back to the general image captioning role and mark the corresponding evidence source as unavailable in the evidence summary. This prevents failed tools from introducing unsupported visual evidence. All prompts and intermediate tool outputs are released with the supplementary code for reproducibility.

D $\text{LLM}_{\textbf{analyst}}$ Supervision and Leakage Control

DIVER uses Llama3-8B as LLManalyst for factual claim extraction, linguistic analysis, and evidence-aware reasoning. To avoid test-set leakage and label shortcuts, all instruction-style supervision is constructed strictly from the official training split. The test split is never used for prompt construction, instruction generation, hyperparameter selection, threshold tuning, or early stopping.

Label-free claim supervision. For claim extraction, the supervision target is generated from raw training text using a label-free claim decomposition prompt. The teacher is asked only to extract atomic, visually or factually verifiable statements

Table 8: Supervision construction and leakage control for LLManalyst.

Objective	Input	Target Construction	Leakage Control
Claim extraction	Raw training text	Label-free atomic claim decomposition generated from text only	No image, test label, dataset identity, or fake/real target is used
Style analysis	Raw training text	Text-only descriptions of sensationalism, emotional exaggeration, and stylistic cues	Describes linguistic phenomena rather than veracity labels
Contradiction analysis	Raw training text and synthetic claim pairs	Entity replacement, number/date perturbation, omission, and contradiction injection	Synthetic labels indicate faithfulness, not news veracity
Evidence-aware reasoning	Training image, verified claim, and extracted visual evidence	Evidence support summary over observable visual facts	No gold rationale or final fake/real label is given to LLManalyst
Validation	Stratified subset of training split	Used for threshold and hyperparameter selection	The official test split is never accessed

from the news text, without access to the ground-truth veracity label, dataset identity, paired image, or any final prediction target. The generated claims are further filtered by semantic faithfulness checks and format constraints. Thus, the claim extractor learns to preserve textual semantics rather than to predict fake/real labels.

Linguistic analysis supervision. For authenticity/style and contradiction analysis, we construct supervision from text-only linguistic cues, such as sensational wording, emotional exaggeration, logical inconsistency, entity mismatch, and temporal inconsistency. These targets describe observable linguistic phenomena rather than dataset labels. In addition, negative claim pairs for the intra-modal judge are synthesized from the training text through entity replacement, number/date perturbation, omission, and contradiction injection. This enables the verifier to learn faithfulness and consistency without using test-set rationales.

Evidence-aware reasoning supervision. For evidence-aware reasoning, the supervision target is an evidence-grounded textual explanation that summarizes whether the extracted visual evidence supports, weakly supports, or fails to support the current claim. Importantly, the LLManalyst is not trained to output the final fake-news label. It only produces intermediate claim and evidence representations, while the final veracity prediction is made by the downstream uncertainty-aware fusion classifier trained with the official training labels.

Unified training and validation protocol. Unless otherwise specified, all datasets follow the same supervision construction pipeline, prompt templates, filtering rules, and decoding settings. We reserve 10% of the official training split as a stratified validation subset for threshold selection, early stopping, and hyperparameter tuning. The official test split remains sealed during all stages of LLManalyst construction and model selection.

E Resource-Matched Fairness Protocol

To address potential resource unfairness, we distinguish between official-performance comparison and resource-matched comparison. Table 9 follows the standard protocol of comparing each baseline under its strongest officially reported or released configuration. In addition, we use the following controls to isolate the contribution of DIVER's dynamic evidence routing from the use of stronger external resources.

Backbone-matched control. We evaluate LLM-only and multimodal reasoning variants under the same 7B/8B-scale backbone whenever possible. This control verifies that the performance gain does not simply come from using a larger language model.

Evidence-matched control. For LLM-based baselines, we provide the same cached visual evidence outputs used by DIVER and require all methods to reason over identical OCR, captioning, dense-captioning, and tagging evidence. Under this setting, the difference lies in how each method selects, organizes, and fuses evidence, rather than in the availability of visual tools.

Budget-matched control. We further compare methods under the same average invocation and token budget. For baselines that exhaustively use visual evidence, we truncate or route evidence under the same invocation budget as DIVER. This evaluates whether DIVER's advantage comes from dynamic evidence allocation rather than more com-

putation.

Open-source-only control. To verify that DIVER is not tied to a proprietary API backend, we instantiate the visual evidence interface with open-source OCR, captioning, dense-captioning, and tagging models. The routing, refinement, and fusion modules remain unchanged. This setting tests whether the proposed dynamic evidence-control mechanism generalizes beyond commercial visual tools.

These controls separate the methodological contribution of DIVER from backbone scale, visual evidence availability, proprietary API capability, and computational budget.

F Additional Routing and Gate Ablations

To further verify that DIVER's improvement comes from dynamic evidence routing rather than merely from additional computation or stronger visual tools, we conduct additional routing and gate ablations. These variants isolate three factors: when Stage 4 is activated, how inter-modal alignment scores are aggregated, and whether the visual evidence backend is necessary.

Routing strategy ablations. We compare DIVER with Random Gate, Always Stage 4, and All Tools. Random Gate activates visual forensics for the same proportion of samples as DIVER but selects samples randomly. Always Stage 4 sends every sample to visual forensics, while All Tools invokes all visual tools for each activated sample. These variants distinguish evidence-necessity estimation from mere activation ratio, exhaustive computation, and indiscriminate tool use.

CLIP aggregation ablations. We compare the proposed lower-bound aggregation with mean and max aggregation over claim-wise CLIP scores. Mean aggregation may hide a single unsupported claim, while max aggregation mainly captures the easiest aligned claim. The lower-bound aggregation is designed to preserve sensitivity to subtle image-text discrepancies.

Resource and backbone ablations. We include an LLM-only fine-tuned baseline using the same LLManalyst backbone but without visual evidence tools. We also include an open-source visual-tool variant to verify that DIVER's gain is not tied to a proprietary API backend.

The results in Table 10 support three observations. First, Random Gate performs worse than DIVER under the same activation ratio, suggesting that the gain does not come from simply invoking visual forensics on more samples, but from correctly estimating evidence necessity. Second, Always Stage 4 and All Tools increase the invocation cost and still underperform DIVER, indicating that exhaustive visual processing introduces redundant or noisy evidence. Third, mean and max CLIP aggregation underperform the proposed lower-bound gate because they can overlook subtle unsupported claims. The LLM-only fine-tuned baseline further shows that the improvement is not solely due to LLManalyst fine-tuning, while the open-source tool variant suggests that DIVER remains effective without relying on a proprietary visual backend.

G Gate Behavior and Robustness Analysis

The Inter-modal Alignment Verification module is designed to regulate reasoning depth rather than directly determine the veracity label. Specifically, $S_{\text{inter}} < \beta$ only indicates insufficient visual grounding or potential cross-modal discrepancy, and therefore activates Evidence-driven Visual Forensics for further verification. Conversely, $S_{\text{inter}} \geq \beta$ indicates that the image provides sufficient coarse semantic support, allowing DIVER to skip expensive fine-grained visual forensics. The final prediction is still made by Uncertainty-Aware Dynamic Fusion rather than by the gate alone.

We use a single validation-selected threshold $\beta = 0.29$ as the default setting across all datasets. This avoids dataset-specific threshold tuning and ensures that the dynamic routing behavior is governed by the same decision criterion. To further examine whether the gate over- or under-activates visual forensics, we report three diagnostic statistics: Stage 4 activation rate, false-trigger rate, and missed-trigger rate.

The diagnostic results show that most visually straightforward samples are filtered before Stage 4, while potentially ambiguous samples are routed to fine-grained evidence extraction. The false-trigger rate reflects the small amount of redundant visual analysis introduced by the conservative gate, whereas the missed-trigger rate measures cases where the lightweight alignment signal fails to activate deeper visual reasoning. Overall, these statistics indicate that the gate provides an effective balance between computational efficiency and evidence coverage.

Formally, the Stage 4 activation rate is defined

Table 9: Resource-matched controls for fair comparison.

Control Setting	Compared Methods	Controlled Resource	Purpose
Official comparison	Released or reported baselines vs. DIVER	Official best configurations	Compare with prior work under standard protocols
Backbone-matched	LLM-only baseline vs. DIVER	Same 7B/8B-scale language backbone	Remove backbone-size advantage
Evidence-matched	LLM-based baselines + cached evidence vs. DIVER	Same OCR/caption/tagging evidence	Test dynamic evidence organization rather than evidence access
Budget-matched	Exhaustive-tool baselines vs. DIVER	Same invocation/token budget	Test whether routing improves cost allocation
Open-source-only	DIVER with open-source visual tools	No proprietary visual API	Verify backend independence and reproducibility

Table 10: Additional routing and gate ablation results for isolating the contribution of dynamic evidence routing. "Invoc." denotes the average amortized model/tool invocations per sample.

Variant	Weibo		Weibo21		GossipCop		Invoc.
	Acc	F1-Fake	Acc	F1-Fake	Acc	F1-Fake	
DIVER	**95.8**	**95.2**	**97.7**	**96.6**	**91.6**	**72.3**	**2.1**
Random Gate	93.1	92.6	94.8	93.9	89.0	68.4	2.1
Always Stage 4	94.0	93.4	95.4	94.6	89.7	69.0	3.9
All Tools	94.5	94.0	96.0	95.1	90.0	69.5	3.4
Mean CLIP Gate	94.4	93.8	96.2	95.2	90.2	69.7	1.8
Max CLIP Gate	93.2	92.7	95.0	94.0	89.4	68.8	1.5
w/o Visual-Verifiability Mask	94.9	94.3	96.7	95.7	90.7	70.4	2.9
Learned Gate	95.5	94.9	97.4	96.2	91.3	71.8	2.2
LLM-only Fine-tuned	91.8	91.2	93.2	92.4	88.3	67.2	1.0
Open-source Tools	94.9	94.4	96.9	95.9	90.8	70.9	2.1

Table 11: Diagnostic statistics of the Inter-modal Alignment Verification gate. Stage 4 Rate denotes the proportion of samples routed to Evidence-driven Visual Forensics. False Trigger denotes unnecessary activation where the non-forensic branch is already correct. Missed Trigger denotes skipped cases that could be corrected by full visual forensics.

Dataset	Stage 4 Rate	False Trigger	Missed Trigger
Weibo	29.6%	8.7%	3.2%
Weibo21	31.4%	9.5%	2.8%
GossipCop	34.1%	11.2%	4.6%

as:

$$r_{\text{stage4}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(S_{\text{inter}}^{(i)} < \beta). \tag{15}$$

The false-trigger rate measures the proportion of samples where Stage 4 is activated although the non-forensic branch already predicts correctly:

$$r_{\text{false}} = \frac{\sum_i \mathbb{I}(S_{\text{inter}}^{(i)} < \beta)\mathbb{I}(\hat{y}_{\text{text}}^{(i)} = y^{(i)})}{\sum_i \mathbb{I}(S_{\text{inter}}^{(i)} < \beta)}. \tag{16}$$

The missed-trigger rate measures skipped samples that are incorrectly predicted by the lightweight branch but can be corrected by the full-forensics branch:

$$r_{\text{miss}} = \frac{\sum_i \mathbb{I}(S_{\text{inter}}^{(i)} \geq \beta)\mathbb{I}(\hat{y}_{\text{text}}^{(i)} \neq y^{(i)})\mathbb{I}(\hat{y}_{\text{full}}^{(i)} = y^{(i)})}{\sum_i \mathbb{I}(S_{\text{inter}}^{(i)} \geq \beta)}. \tag{17}$$

H Efficiency Measurement Protocol

To ensure a fair efficiency comparison, we use the same runtime protocol for all LLM-based methods evaluated in Table 3 and Table 4. Latency is measured as the wall-clock processing time for one

Table 12: Efficiency measurement protocol used for latency, model/tool invocation, and token statistics.

Item	Protocol
Latency	Wall-clock processing time per sample, including local LLM reasoning, activated visual tools, post-processing, and final fusion. For API-based tools, unstable remote-service queuing time is excluded.
Hardware	Local neural modules are evaluated on the same NVIDIA Tesla A100 GPU.
Batching	All methods are evaluated with batch size 1 to reflect online fake-news verification.
Model/Tool Invocations	Each structured LLManalyst generation or activated image-to-text evidence request is counted as one invocation. The three linguistic objectives in Stage 1 are produced by a single structured LLM-analyst invocation rather than three independent calls. Skipped visual tools contribute zero invocations.
Token Count	Counts all text input/output tokens from LLM reasoning, prompts, self-correction, evidence summaries, and final decision generation.
Visual Tools	OCR, image captioning, dense captioning, and image tagging use the same image-to-text backend, model version, prompt template, and decoding setting across datasets.
Averaging	All reported latency, invocation, and token statistics are averaged over the evaluated test samples.

news sample, starting from the raw text-image input and ending at the final veracity prediction. This includes local LLM reasoning, activated visual evidence extraction, post-processing, and final fusion. For API-based visual tools, we exclude only unstable remote-service queuing time, while keeping the same API environment and decoding setting for all compared methods. We use batch size 1 for all methods to simulate online verification scenarios.

This protocol reflects the amortized cost of DIVER under dynamic routing. For visually consistent samples, the Inter-modal Alignment Verification stage skips Evidence-driven Visual Forensics, reducing both model/tool invocations and token usage. For ambiguous samples, DIVER spends additional computation on fine-grained evidence extraction, but this cost is amortized over the full test set. Therefore, the reported efficiency numbers measure the practical cost of the complete dynamic pipeline rather than the cost of a single isolated module.

In particular, although Linguistic Investigation contains factual claim extraction, authenticity/style analysis, and contradiction analysis, these objectives are implemented as a single structured LLManalyst generation with a unified output format. Thus, they are counted as one model invocation. Additional invocations are incurred only when self-correction or visual evidence extraction is activated. This definition avoids conflating functional objectives with separate model calls and makes the reported overhead consistent with the actual dynamic execution process.

Table 13: Statistics of the evaluated datasets.

Dataset	Total	Real	Fake	Split	
				Train	Test
Weibo	9,528	4,745	4,783	7,532	1,996
Weibo21	9,127	4,640	4,487	7,301	1,826
GossipCop	12,840	10,259	2,581	10,010	2,830

I Dataset Statistics and Split Details

We provide detailed statistics of the evaluated datasets in Table 13. The three benchmarks cover different linguistic and topical settings: Weibo and Weibo21 are Chinese multimodal fake news datasets, while GossipCop focuses on English entertainment news. This combination enables evaluation across both language and domain variations.

As shown in Table 13, Weibo and Weibo21 are relatively balanced between real and fake news, whereas GossipCop exhibits a stronger class imbalance with substantially more real samples than fake ones. Therefore, besides Accuracy, we report F1-Fake, F1-Real, and AUC in the main experiments to provide a more comprehensive evaluation of both fake-news detection and real-news recognition. During training, all supervision is constructed only from the official training split, and the test split is used solely for final evaluation.

J Conceptual Comparison with Existing Paradigms

To further clarify the methodological difference, Table 14 compares DIVER with representative multimodal fake news detection paradigms. Here, "Visual Evidence" refers to explicit claim-conditioned evidence, such as OCR, dense captions, and image tags, rather than generic visual embeddings.

Table 14: Conceptual comparison between DIVER and existing multimodal fake news detection paradigms.

Paradigm	Claim-level	Dynamic Gate	Visual Evidence	Masked Fusion
Static multimodal fusion	✗	✗	✗	✗
LLM-based reasoning	✓	✗	✗	✗
Tool-augmented verification	✓	✗	✓	✗
DIVER	✓	✓	✓	✓

K Cross-Domain Generalization Analysis

To further evaluate whether DIVER generalizes beyond the standard in-domain train/test splits, we conduct additional cross-domain transfer experiments. Different from the main experiments, where models are trained and evaluated on the standard split of the same dataset, the cross-domain setting trains the model on a source dataset and directly evaluates it on a different target dataset. This protocol is more challenging because the training and test sets contain different event distributions, topic patterns, and dataset-specific biases.

Specifically, we consider two source-only transfer settings: Weibo $\rightarrow$ Weibo21 and Weibo21 $\rightarrow$ Weibo. For each setting, all trainable components are optimized only on the source training split, including the intra-modal consistency verifier, the final fusion classifier, and the LLM-based linguistic analyst. Hyperparameters such as the inter-modal gating threshold β, the refutation threshold γ, and the maximum correction round τ are selected using only the source validation split. The target dataset is used exclusively for testing, and no target-domain labels, validation samples, or rationales are used during training, threshold selection, or model selection.

This evaluation examines whether the proposed dynamic evidence routing mechanism remains robust under dataset-level distribution shifts. In particular, models that mainly rely on dataset-specific correlations or static multimodal fusion patterns often degrade substantially when transferred to a different dataset. In contrast, DIVER explicitly decomposes textual claims, estimates the necessity of visual evidence through inter-modal alignment, and selectively activates fine-grained visual forensics only for potentially inconsistent samples. This design helps reduce over-reliance on source-domain artifacts and improves robustness on unseen target distributions.

The results are reported in Table 15. Compared with static multimodal fusion methods and LLM-based reasoning baselines, DIVER achieves stronger cross-domain performance in both transfer directions. These results indicate that the proposed alignment-guided evidence routing strategy is not merely fitted to in-domain train/test splits, but can better preserve multimodal reasoning ability when the event distribution changes.

L Parameter Sensitivity Analysis

To provide a more complete evaluation of hyperparameter robustness, we report sensitivity results on F1-Real, Accuracy, F1-Fake, and AUC. Specifically, we vary four key hyperparameters: the inter-modal gating threshold β, LLM temperature, the visual-evidence scaling coefficient λ, and the number of extracted factual claims k. The results are shown in Fig. 4, Fig. 5, Fig. 6, and Fig. 7. Overall, DIVER remains stable under moderate perturbations of these hyperparameters.

F1-Real sensitivity. As shown in Fig. 4, DIVER shows stable F1-Real trends across Weibo, Weibo21, and GossipCop. The inter-modal gating threshold β is validation-selected rather than learned by back-propagation, and the best or near-best validation performance is achieved around $\beta = 0.29$. Similarly, the model performs consistently well around temperature $= 0.40$ and visual-evidence scaling coefficient $\lambda = 0.40$. For claim decomposition, performance generally peaks at $k = 4$ and slightly saturates or degrades when more claims are extracted, reflecting a coverage–noise trade-off. Therefore, we set $\beta = 0.29$, temperature $= 0.40$, $\lambda = 0.40$, and $k = 4$ as the default configuration.

Accuracy sensitivity. As shown in Fig. 5, Accuracy changes smoothly across all three datasets. The best or near-best performance is generally obtained around the validation-selected threshold $\beta = 0.29$, temperature $= 0.40$, and visual-evidence scaling coefficient $\lambda = 0.40$. For claim decomposition, Accuracy improves as more factual claims are extracted and usually peaks around $k = 4$, after which the gains saturate or slightly decline. This confirms that the default setting provides a good balance between evidence coverage and noise control.

F1-Fake sensitivity. Fig. 6 further evaluates the stability of DIVER on fake-news detection. Although F1-Fake is more sensitive than Accuracy, especially on GossipCop due to its stronger class imbalance, the overall curves remain stable and do not exhibit abrupt collapse under moderate hyperparameter changes. The results show that the

Method	Weibo → Weibo21				Weibo21 → Weibo			
	Acc	F1-Fake	F1-Real	AUC	Acc	F1-Fake	F1-Real	AUC
FND-CLIP	87.6	87.2	88.1	92.8	86.9	87.3	86.5	92.1
MIMoE-FND	89.4	89.0	89.8	94.1	88.7	89.2	88.3	93.7
KEN	88.9	88.6	89.3	93.6	89.1	89.5	88.7	94.0
LIFE	88.2	88.7	87.8	93.2	87.6	88.1	87.0	92.8
INSIDE	86.5	85.9	87.1	91.5	85.8	86.2	85.4	90.9
FactGuard	85.7	85.2	86.1	90.8	84.9	85.5	84.3	90.1
DIVER	**92.1**	**91.6**	**92.5**	**96.0**	**91.8**	**92.2**	**91.4**	**95.7**

Table 15: Cross-domain generalization results under the source-only transfer protocol. Models are trained only on the source dataset and directly evaluated on the target dataset. No target-domain labels, validation samples, or rationales are used for training, threshold selection, or model selection. All results are reported as percentages (%), averaged over three random seeds.

proposed dynamic gating and evidence-driven refinement do not merely improve real-news recognition, but also preserve robust fake-news detection performance.

AUC sensitivity. As shown in Fig. 7, AUC remains consistently high across different hyperparameter settings, indicating that DIVER preserves strong ranking robustness even when the decision threshold or evidence granularity changes. The best AUC values are also obtained near the default configuration, which supports the choice of $\beta = 0.29$, temperature $= 0.40$, visual-evidence scaling coefficient $\lambda = 0.40$, and $k = 4$. Together with the F1-Real, Accuracy, and F1-Fake results, these curves demonstrate that the selected hyperparameters are not over-specialized to a single metric.

Overall, the additional metrics lead to the same conclusion: DIVER is not highly sensitive to small hyperparameter perturbations. The validation-selected configuration achieves strong or near-optimal results across F1-Real, Accuracy, F1-Fake, and AUC, supporting the robustness of the alignment-gated visual forensics and uncertainty-aware fusion modules.

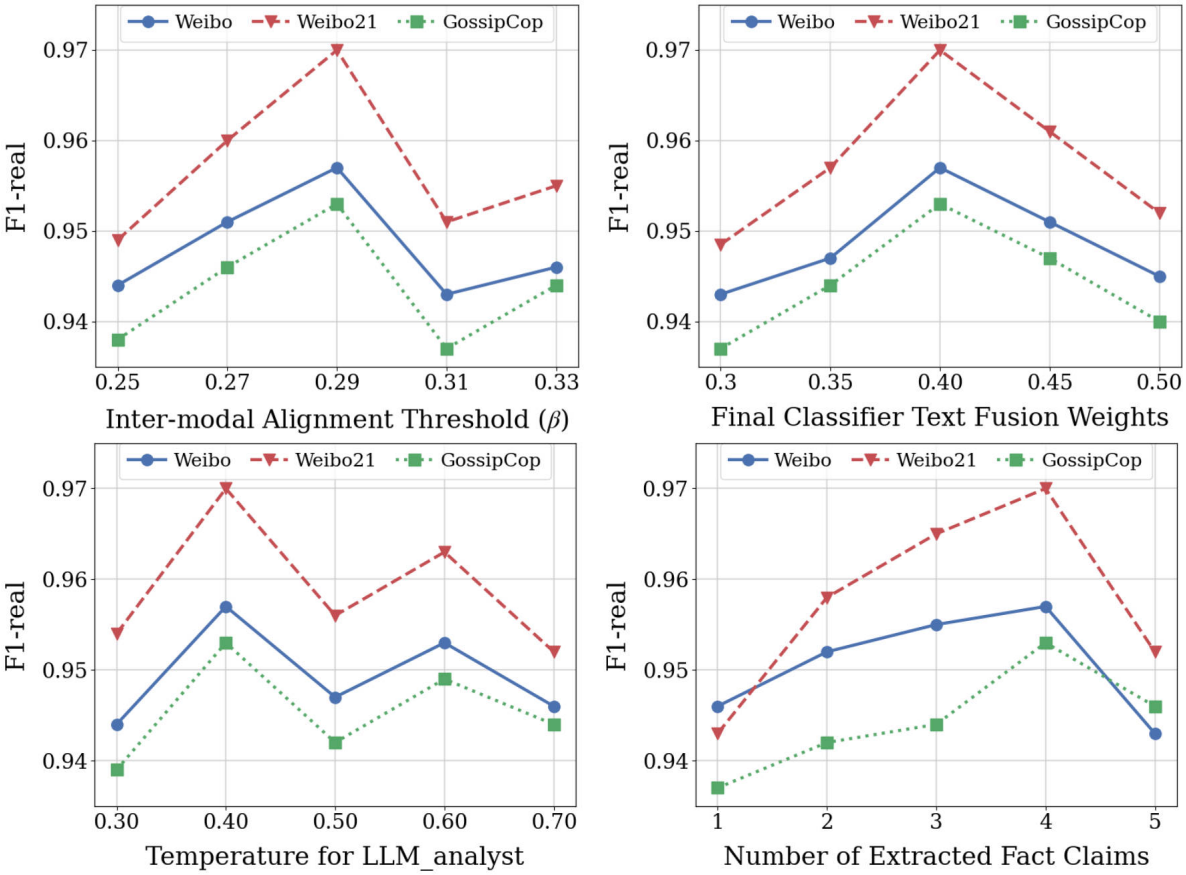

Figure 4: Parameter sensitivity analysis on F1-Real across Weibo, Weibo21, and GossipCop.

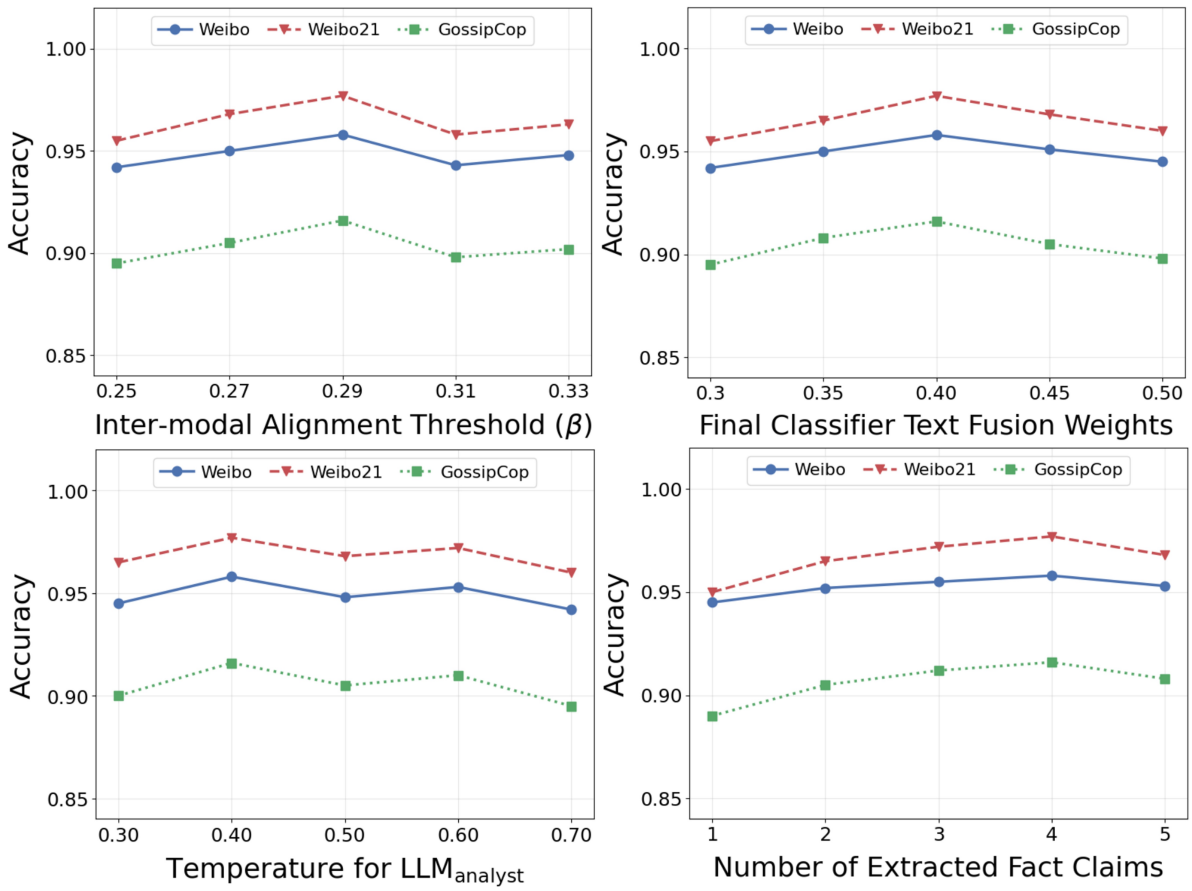

Figure 5: Additional parameter sensitivity analysis on Accuracy across Weibo, Weibo21, and GossipCop.

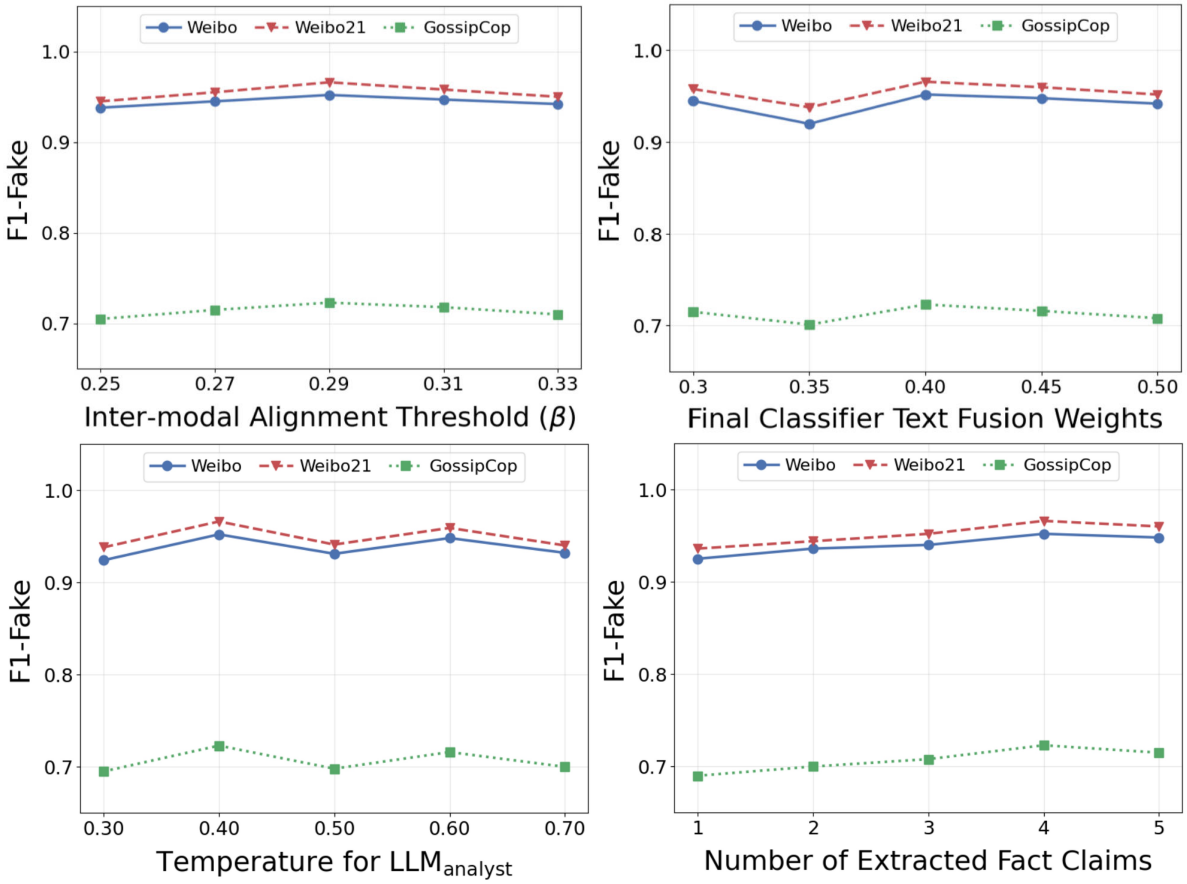

Figure 6: Additional parameter sensitivity analysis on F1-Fake across Weibo, Weibo21, and GossipCop.

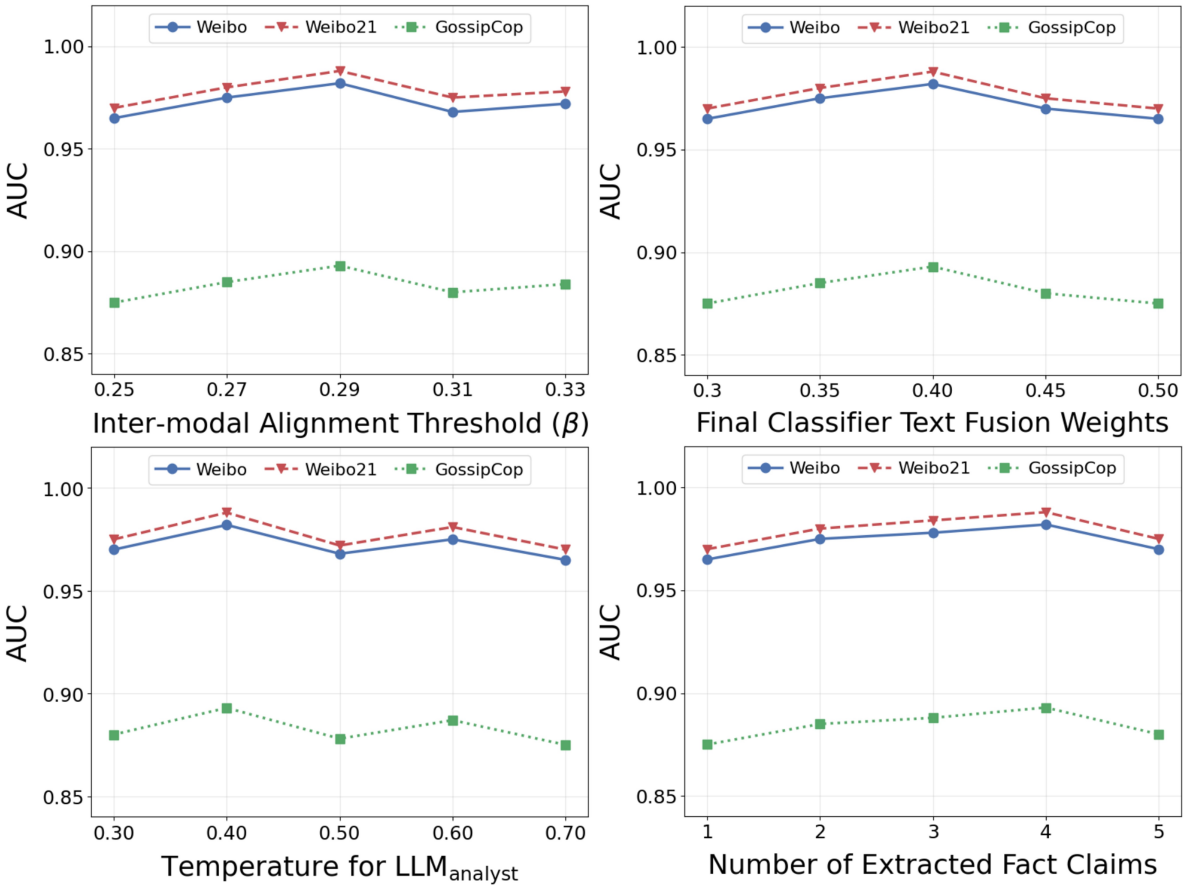

Figure 7: Additional parameter sensitivity analysis on AUC across Weibo, Weibo21, and GossipCop.