GLED-NET: A GLOBAL–LOCAL EVIDENCE DECOMPOSITION NETWORK FOR MULTIMODAL FAKE NEWS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fake news detection requires reasoning about both the overall consistency of an image–text pair and a compact set of textual cues that may carry discriminative information. We propose GLED-Net, a CLIP-based global–local evidence decomposition framework that organizes these views into a global narrative stream and a local cue stream. The global stream models the correspondence between the complete post and its image, whereas the local stream uses automatically extracted keywords to construct a complementary representation. Branch-level supervision encourages both streams to remain predictive, and a learned gate combines their outputs. Experiments on Weibo-16 and FNDC show strong performance relative to the evaluated supervised and zero-shot baselines. Same-backbone controls, component ablations, sensitivity analysis, and repeated runs distinguish the effect of the pretrained encoder from that of the proposed evidence organization. The results also show that the benefit of heuristic keyword cues and adaptive fusion is dataset dependent, which motivates more adaptive and verifiable cue selection.

1 INTRODUCTION

Multimodal fake news detection aims to determine whether a news post composed of an image and its accompanying text is real or fake Shu et al. (2017; 2019); Jin et al. (2017a); Wang et al. (2018); Chen et al. (2022); Qi et al. (2023). Given an image-text pair (I, T) from social media, the model is expected to output a veracity label $y \in \{0, 1\}$, where 0 and 1 denote real and fake news, respectively. This task is increasingly important because image-text misinformation is often more persuasive and more rapidly disseminated than unimodal content Jin et al. (2017a); Wang et al. (2018); Luo et al. (2021); Qi et al. (2023); Bu et al. (2024). Compared with text-only rumor and fake-news detection Ma et al. (2017; 2018); Bian et al. (2020); Yang et al. (2020), multimodal fake news detection requires the model to jointly reason over visual semantics, textual claims, and their cross-modal consistency, making it substantially more challenging in realistic social-media environments.

Despite recent progress, existing methods still struggle to robustly identify deceptive multimodal fake news. Recent multimodal detectors have explored direct feature fusion, event-invariant representation learning, knowledge-aware graphs, co-attention fusion, ambiguity modeling, and mixture-of-experts reasoning Jin et al. (2017a); Wang et al. (2018; 2020); Wu et al. (2021); Chen et al. (2022); Zheng et al. (2022); Ma et al. (2024); Liu et al. (2025), yet several difficulties remain unresolved. First, fake news is often *narratively coherent* at the paragraph level: the text may sound fluent and plausible, while the falsehood is only exposed through subtle image-text inconsistency. Second, the critical evidence is usually *sparsely distributed* in a few salient words, entities, or attributes rather than the entire text. Third, social media data are inherently noisy, containing informal language, emotional expressions, loosely related images, or even reused images, which make coarse-grained multimodal fusion prone to learning superficial correlations instead of reliable evidence Qi et al. (2023); Tong et al. (2024); Bu et al. (2024); Qiao et al. (2025). As a result, directly collapsing the whole image-text pair into a single representation may blur the distinction between genuine visual support and accidental semantic overlap.

To address these challenges, we study two complementary views of a multimodal post: a *global narrative view*, which captures the correspondence between the complete post and its image, and a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

local cue view, which summarizes a small set of salient words, entities, or event descriptors. This distinction is a task-oriented modeling hypothesis rather than a claim that each selected keyword constitutes causal or human-verified evidence. Based on this hypothesis, we propose **GLED-Net**, a **G**lobal–**L**ocal **E**vidence **D**ecomposition Network built on a pretrained CLIP Radford et al. (2021); Yang et al. (2022). GLED-Net learns branch-specific representations and combines them with a lightweight gate. Feature gating, discrepancy features, auxiliary losses, and gated fusion are established neural components; our contribution lies in their task-specific organization into separately supervised global and local views for fake news detection.

As shown in Figure 1, GLED-Net contains a CLIP-based encoder, a global narrative stream, a local keyword-conditioned stream, and a fusion module. The design offers a direct way to organize heterogeneous post-level signals without external retrieval or large-model reasoning. Component ablations below separately evaluate the pretrained backbone, both streams, branch supervision, and adaptive fusion. The main contributions of this work are summarized as follows:

- We formulate multimodal fake news detection using two complementary views: full-post narrative consistency and compact keyword-conditioned cues.
- We present a task-oriented CLIP-based architecture that learns the two views in separate branches, applies branch-level supervision, and combines them through a learned fusion function.
- We provide same-backbone controls, component ablations, sensitivity analyses, and five-seed results that separate the contribution of CLIP adaptation from the proposed evidence organization and expose its limits.

2 Related Work

Multimodal Fake News Detection. Multimodal fake news detection has evolved from early feature fusion paradigms to more structured reasoning frameworks. Early representative studies focused on direct text-image fusion and event-invariant representation learning Jin et al. (2017a); Wang et al. (2018). Later methods introduced knowledge-aware graph modeling, co-attention fusion, ambiguity-aware learning, and multimodal-social joint reasoning to strengthen cross-modal interaction and robustness Wang et al. (2020); Wu et al. (2021); Chen et al. (2022); Zheng et al. (2022). In parallel, recent benchmarks and problem settings such as NewsCLIPpings and FakeSV have broadened the scope of research from traditional image-text posts to out-of-context misinformation and short-video fake news Luo et al. (2021); Qi et al. (2023); Bu et al. (2024). More recent studies have further emphasized event-level inconsistency, multi-view learning, domain generalization, and cross-modal credibility modeling. For example, Event-Radar Ma et al. (2024) captures event-driven multimodal inconsistency through graph-based multi-view reasoning and explicitly models view credibility under noisy samples. ViKI Li et al. (2024) improves multimodal disinformation detection by learning aligned vision-language embeddings and cross-modal knowledge interaction. MMDFND Tong et al. (2024) studies multi-domain multimodal fake news detection, while recent journal work also explores stronger cross-modal content correlation modeling Qiao et al. (2025). More recently, researchers have explored stronger open-world reasoning ability by incorporating large vision-language models, synthetic supervision, or explainable out-of-context reasoning. FKA-Owl Liu et al. (2024) augments LVLMs with forgery-specific knowledge, Zeng *et al.* Zeng et al. (2024) study how synthetic data can improve real-world multimodal misinformation detection with multimodal LLMs, Tahmasebi *et al.* Tahmasebi et al. (2024) analyze multimodal misinformation detection with large vision-language models, and Sniffer Qi et al. (2024) pushes explainable out-of-context misinformation detection with MLLMs. MIMoE-FND Liu et al. (2025) further highlights the importance of explicit modality interaction in multimodal fake news detection. Although these methods have significantly advanced the field, most of them still focus on event/entity reasoning, domain adaptation, external knowledge augmentation, or data scaling. In contrast, our work emphasizes a different but underexplored issue: the heterogeneity of evidence inside a single multimodal post. We argue that the full narrative and sparse discriminative cues should not be treated as a single undifferentiated signal. GLED-Net therefore decomposes multimodal evidence into narrative-level and cue-level branches before fusion, which is particularly suitable for noisy social-media fake news.

Vision-Language Pre-training for Multimodal Reasoning. Large-scale vision-language pre-training has become a strong foundation for multimodal understanding. Earlier pretrained vision-language models such as ViLBERT Lu et al. (2019), LXMERT Tan & Bansal (2019), and UNITER

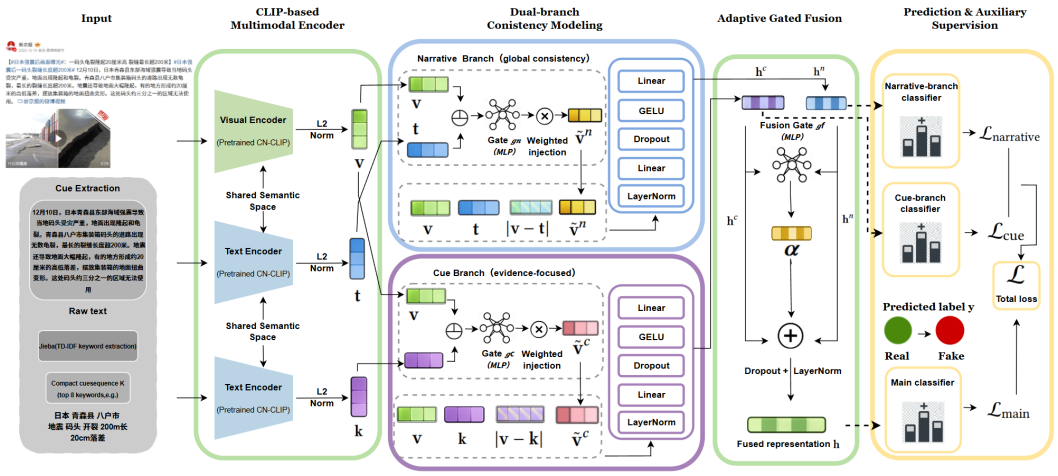

Figure 1: GLED-Net architecture. A Chinese CLIP encoder maps image I, full text T, and TF–IDF cue sequence K into a shared space. The global stream constructs $\mathbf{h}^n$ from the image and full text, while the local stream constructs $\mathbf{h}^c$ from the image and compact cue sequence. A learned gate combines both representations for classification, and branch classifiers provide auxiliary supervision. The cue sequence is automatically generated and is not treated as human-verified evidence.

Chen et al. (2020) demonstrate that aligned image-text representation learning substantially improves downstream multimodal reasoning. More recent contrastive and generative pre-training paradigms, including CLIP Radford et al. (2021), ALBEF Li et al. (2021), BLIP Li et al. (2022a), and BLIP-2 Li et al. (2023), further show that large-scale image-text pre-training offers strong transferability and efficiency. Chinese CLIP Yang et al. (2022) extends this paradigm to Chinese, offering aligned image-text representations that are well-suited to Chinese social-media content. These advances suggest that pretrained vision-language models provide a much stronger initialization than training multimodal detectors from scratch. However, directly applying global image-text representations is often insufficient for misinformation detection. Fake news frequently depends on a subtle cross-modal mismatch rather than coarse semantic similarity. Therefore, the key challenge is not only to obtain aligned multimodal features, but also to reorganize them into evidence structures that are discriminative for veracity reasoning. Our method follows the vision-language pre-training paradigm, but goes one step further by explicitly decomposing the aligned features into narrative evidence and cue evidence, making the learned representation more suitable for fake news detection instead of generic image-text matching.

Fine-grained Cross-modal Alignment and Evidence Reasoning. Another relevant line of work investigates fine-grained cross-modal reasoning and grounding. MDETR Kamath et al. (2021), RegionCLIP Zhong et al. (2022), and GLIP Li et al. (2022b) show that fine-grained language-image alignment is crucial when the task depends on localized semantic evidence. In multimodal misinformation analysis, HAMMER Shao et al. (2023) further demonstrates the importance of fine-grained interaction by jointly detecting and grounding manipulated image regions and text tokens. Sniffer Qi et al. (2024) also shows that explainable out-of-context misinformation detection benefits from more detailed cross-modal discrepancy discovery rather than pure global consistency estimation. Our work is related to this direction but differs in both objective and formulation. Instead of performing explicit grounding supervision on manipulated regions or tokens, GLED-Net focuses on veracity prediction under weak supervision and models fine-grained evidence through a cue-guided branch. This design preserves the efficiency of a classification-oriented detector while still benefiting from clue-level reasoning. In this sense, our framework can be viewed as a lightweight evidence decomposition strategy between coarse narrative semantics and compact local cues, bridging the gap between global fusion detectors and fully grounded manipulation reasoning models.

3 METHOD

3.1 PROBLEM FORMULATION

Given a multimodal news post $\mathcal{X} = \{I, T\}$, where I denotes the image and T denotes the associated text, our goal is to predict its veracity label $y \in \{0, 1\}$, where 0 and 1 indicate real and fake news, respectively. Beyond the full post text, we further derive a compact cue sequence K from T to highlight potentially discriminative evidence. The key motivation is that multimodal fake news often contains two complementary forms of evidence: 1) *narrative-level evidence*, which reflects the global semantic consistency between the image and the full textual claim, and 2) *cue-level evidence*, which is concentrated in a few salient keywords, entities, or attributes and is often more effective for exposing subtle deception. To model these two forms of evidence explicitly, we propose **GLED-Net**, a CLIP-based **G**lobal-**L**ocal **E**vidence **D**ecomposition Network that separately learns a narrative branch and a cue branch, and then adaptively fuses them under auxiliary supervision.

3.2 OVERVIEW

As illustrated in Figure 1, GLED-Net consists of four modules. A CLIP encoder produces aligned image and text representations. A narrative branch processes the image and complete post, and a cue branch processes the image and TF–IDF keyword sequence. A learned fusion module combines the two branch representations, while auxiliary classifiers encourage both branches to remain predictive. This is a task-oriented post-level architecture for organizing global narrative and local cue representations. It does not rely on external knowledge retrieval or large-model reasoning, although its computational cost is still dominated by the pretrained Chinese-CLIP backbone.

3.3 CLIP-BASED MULTIMODAL ENCODING

Recent vision-language pre-training methods have shown strong transferability for downstream multimodal reasoning Radford et al. (2021); Li et al. (2023); Zhong et al. (2022). Following this paradigm, we instantiate GLED-Net with a pretrained Chinese CLIP Yang et al. (2022), which is well suited to Chinese social-media content.

For each sample, the visual encoder $f_v(\cdot)$ and the text encoder $f_t(\cdot)$ produce a global visual feature, a narrative text feature, and a cue text feature:

$$\mathbf{v} = \mathrm{norm}(f_v(I)), \quad \mathbf{t} = \mathrm{norm}(f_t(T)), \quad \mathbf{k} = \mathrm{norm}(f_t(K)), \tag{1}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. These normalized features lie in the same CLIP-aligned semantic space, which facilitates both cross-modal matching and discrepancy modeling.

Local Cue Extraction. To construct the cue sequence K, we first clean the raw post text by removing URLs, user mentions, and hashtag markers. We then use the TF–IDF keyword extractor in Jieba[1] to select the top-K terms, with $K = 4$ by default. If no keyword is returned, we use the cleaned raw text as a fallback. These automatically constructed terms are weak cues rather than human-annotated or causal evidence. Accordingly, the local stream should be interpreted as a compact complementary text view, not as an explicit grounding or explanation module.

3.4 DUAL-BRANCH CONSISTENCY MODELING

Global Narrative-level Branch for Global Consistency Modeling. The narrative branch captures the overall semantic relation between the complete post text and the image. Instead of directly concatenating $\mathbf{v}$ and $\mathbf{t}$, we first use a learnable gate to inject narrative semantics into the visual representation as follows:

$$\tilde{\mathbf{v}}^n = \mathbf{v} + g_n([\mathbf{v}; \mathbf{t}]) \odot \mathbf{t}, \tag{2}$$

where $[\cdot;\cdot]$ denotes feature concatenation, $\odot$ denotes element-wise multiplication, and $g_n(\cdot)$ is a two-layer fully-connected network with a sigmoid activation. This operation adaptively controls how

[1] https://github.com/fxsjy/jieba

much narrative information should refine the visual feature for each sample. Then, we can construct the narrative-branch representation $\mathbf{h}^n$ as follows:

$$\mathbf{h}^n = \phi_n([\mathbf{v}; \mathbf{t}; |\mathbf{v} - \mathbf{t}|; \tilde{\mathbf{v}}^n]), \tag{3}$$

where $|\mathbf{v} - \mathbf{t}|$ explicitly characterizes cross-modal discrepancy and $\phi_n(\cdot)$ is a lightweight feed-forward projection block. In the implementation, $\phi_n(\cdot)$ consists of a linear projection layer, GELU activation, dropout, another linear projection layer, and layer normalization. This branch is expected to capture whether the image globally supports the claim expressed in the full news text.

Local Cue-level Branch for Evidence-focused Verification. Although the full post preserves context, it may also contain redundant or informal expressions. We therefore construct a second representation from the compact cue sequence K. Similar to the narrative branch, we obtain a cue-conditioned visual representation as follows:

$$\tilde{\mathbf{v}}^c = \mathbf{v} + g_c([\mathbf{v}; \mathbf{k}]) \odot \mathbf{k}, \tag{4}$$

where $g_c(\cdot)$ is another learnable gating function with the same structure as $g_n(\cdot)$. Then, the cue-branch representation $\mathbf{h}^c$ is defined as follows:

$$\mathbf{h}^c = \phi_c([\mathbf{v}; \mathbf{k}; |\mathbf{v} - \mathbf{k}|; \tilde{\mathbf{v}}^c]), \tag{5}$$

where $\phi_c(\cdot)$ shares the same feed-forward architecture as $\phi_n(\cdot)$. This branch supplies a compact keyword-conditioned view of the post. We do not assume that the selected terms are faithful evidence; the component and sensitivity analyses evaluate whether this additional view is useful under each dataset.

3.5 Adaptive Gated Fusion

The global narrative-level branch and local cue-level branch are complementary. The global branch emphasizes holistic claim-level semantics, while the local branch highlights condensed evidence. To integrate them adaptively, we compute a fusion gate and obtain the fused representation as follows:

$$\mathbf{h} = \boldsymbol{\alpha} \odot \mathbf{h}^n + (1 - \boldsymbol{\alpha}) \odot \mathbf{h}^c, \quad \boldsymbol{\alpha} = g_f([\mathbf{h}^n; \mathbf{h}^c]), \tag{6}$$

where $g_f(\cdot)$ is a two-layer MLP with GELU and sigmoid activations. The fused representation $\mathbf{h}$ is regularized by dropout and layer normalization before classification. The gate provides sample-dependent weights, but these weights are not interpreted as calibrated reliability or faithful explanations. We compare this learned gate with fixed mean fusion in the component ablation.

3.6 Optimization Objective

To encourage both global narrative-level and local cue-level discriminability while maintaining robust final fusion, we attach one classifier to each branch and one classifier to the fused representation. Let $\mathcal{L}_{\text{main}}$, $\mathcal{L}_{\text{global}}$, and $\mathcal{L}_{\text{local}}$ denote the cross-entropy losses of the fused prediction, the narrative-branch prediction, and the cue-branch prediction, respectively. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \frac{\lambda}{2} \left(\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} \right), \tag{7}$$

where λ controls the contribution of the auxiliary branch supervision. This objective stabilizes optimization by ensuring that each branch remains individually predictive instead of collapsing into a weak auxiliary pathway. In our released implementation, λ is set to 0.2 by default.

Discussion. Our design is motivated by recent advances in multimodal fake news detection and multimodal manipulation analysis Ma et al. (2024); Liu et al. (2024); Shao et al. (2023); Li et al. (2024). Different from methods that introduce external knowledge graphs, synthetic supervision, or heavy large-model reasoning, GLED-Net emphasizes a practically deployable formulation with three desirable properties: 1) strong initialization from a pretrained vision-language model, 2) explicit decomposition into narrative-level and cue-level evidence, and 3) adaptive fusion with auxiliary branch regularization. These properties make the model well-suited to social-media scenarios, where posts are short, noisy, and highly dependent on subtle image-text inconsistency.

Table 1: Data statistics of Weibo-16 and FNDC.

	Weibo-16			FNDC		
Split	Real	Fake	Total	Real	Fake	Total
Training set	4,100	4,100	8,200	4,500	4,500	9,000
Test set	1,145	1,145	2,290	500	500	1,000
Total	5,245	5,245	10,490	5,000	5,000	10,000

4 EXPERIMENTS

4.1 SETTINGS

Datasets. To comprehensively evaluate the performance of GLED-Net, we conduct experiments on two multimodal fake news detection datasets, Weibo-16 Jin et al. (2017b) and FNDC KOTO (2019). The Weibo-16 dataset is a widely used benchmark for multimodal fake news detection, originally constructed from posts collected on Weibo, a major Chinese social media platform Jin et al. (2017b). It contains diverse real-world misinformation cases spanning topics such as social events, public health, and entertainment. Following data cleaning and preprocessing, the version used in our experiments contains 10,490 image-text pairs in total. As shown in Table 1, the training set consists of 8,200 samples, including 4,100 real and 4,100 fake news pieces, while the test set contains 2,290 samples, with 1,145 real and 1,145 fake samples; both splits are balanced at a 1:1 real/fake ratio. The FNDC dataset was released by Zhiyuan and the Institute of Computing Technology, Chinese Academy of Sciences, and was used in the 2019 Internet Fake News Detection Competition[2]. We match each CSV record to locally available images using its `piclist` field rather than using `id` as the primary image-matching key. When a post contains multiple matched images, each matched image forms a separate image–text pair with the same text and label. After discarding records without a matched image, we balance the two classes by subsampling the larger class to the size of the smaller class. The final FNDC subset contains 10,000 samples: using seed 42, we sample 5,000 real and 5,000 fake pairs and split them into 9,000 training samples (4,500 per class) and 1,000 test samples (500 per class). During model training, 10% of the 9,000-sample training set is further reserved for validation. This procedure yields the statistics in Table 1.

Evaluation Metrics. Considering that the fake news detection task is a binary classification task (real/fake), we deploy multiple classification metrics to comprehensively evaluate the model's performance, including Accuracy (Acc), Precision (Prec), Recall (Rec), and F1 score.

Baselines. We compare our model with the following baselines: **Qwen3-VL-4B-Instruct** Bai et al. (2025), **DeepSeek-VL2-Tiny** Wu et al. (2024), **FND-Bootstrap** Ying et al. (2023), **FND-CLIP** Zhou et al. (2023), and **MIMoE-FND** Liu et al. (2025).

Implementation Details. All experiments were conducted on a workstation equipped with one NVIDIA RTX 4090D GPU and one NVIDIA RTX 3090 GPU. GLED-Net was implemented in PyTorch and initialized with `iic/multi-modal_clip-vit-base-patch16_zh`. We use AdamW with learning rate 1×10^{-5} for CLIP fine-tuning, weight decay 1×10^{-4}, and cosine learning-rate annealing. Unless otherwise stated, all GLED-Net runs use 5 epochs, batch size 16, seed 42, keyword budget $K = 4$, and auxiliary-loss weight $\lambda = 0.2$. For image processing, all input images are resized to 224×224. In the CLIP setting, images are normalized using the default CLIP mean and standard deviation. During training, we apply random horizontal flipping as the only data augmentation strategy, while no augmentation is used in validation or testing. For text processing, noisy tokens such as URLs, user mentions, and redundant whitespace are removed before tokenization. We further extract keyword-aware textual evidence using a Chinese keyword extractor, where the top-4 keywords are retained for the auxiliary keyword branch. In the CLIP-based configuration, both the raw text and the keyword sequence are truncated to a maximum length of 512 tokens. The dropout rate is set to 0.1, and the auxiliary branch loss weight is set to 0.2. Following common practice in fake news detection, the best checkpoint is selected according to the validation F1 score. For datasets without an official validation split, we reserve 10% of the training split for validation and use the remaining portion for model training. In addition to supervised baselines, the

[2]https://www.heywhale.com/mw/dataset/5e096a422823a10036b10a81

Table 2: Comparison results on the Weibo-16 dataset and the FNDC dataset.

	Weibo-16				FNDC			
Method	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Qwen3-VL-4B-Instruct	60.96%	56.25%	**98.69%**	71.66%	22.75%	24.32%	75.47%	36.78%
DeepSeek-VL2-Tiny	54.89%	66.67%	19.56%	30.25%	43.50%	30.66%	10.75%	15.92%
FND-Bootstrap†	89.87%	91.69%	87.69%	89.64%	96.40%	97.15%	95.60%	96.37%
FND-CLIP	87.55%	92.74%	81.48%	86.75%	97.80%	98.78%	96.80%	97.78%
MIMoE-FND	88.91%	92.47%	84.72%	88.42%	97.80%	**99.38%**	96.20%	97.76%
GLED-Net (Ours)	**94.06%**	**95.41%**	92.58%	**93.97%**	**98.30%**	98.99%	**97.60%**	**98.29%**

large vision-language models Qwen3-VL-4B-Instruct and DeepSeek-VL2-Tiny are evaluated in a zero-shot manner under a unified prompting protocol, where both the input image and the associated text are provided to the model, and the model is required to output a structured JSON prediction indicating whether the sample is `real` or `fake`.

4.2 Results and Analysis

Main Results. Table 2 reports point estimates on Weibo-16 and FNDC. GLED-Net obtains the highest F1 among the evaluated methods on both datasets. We avoid claims of statistical significance because the baselines were not repeated under the same random seeds; optimization stability is analyzed separately in Appendix C.

On Weibo-16, GLED-Net achieves 94.06% Accuracy, 95.41% Precision, 92.58% Recall, and 93.97% F1, achieving better performance under the evaluated setting than the strongest baseline FND-Bootstrap. It also performs better than FND-CLIP and MIMoE-FND in the reported experiments. Since Weibo-16 contains noisy and diverse social-media content, these gains suggest that the proposed dual-branch design is useful for handling subtle image-text inconsistency in this benchmark.

On FNDC, the full model rerun in the common ablation environment achieves 98.30% Accuracy, 98.99% Precision, 97.60% Recall, and 98.29% F1. This is a strong balance of precision and recall under the evaluated protocol. The component analysis below nevertheless shows that fixed mean fusion is slightly better on this dataset, so the FNDC result should not be interpreted as evidence that every proposed component is uniformly beneficial.

For FND-Bootstrap, the MAE checkpoint was unavailable in the new environment. Therefore, the MAE image encoder was randomly initialized using the `--no-pretrained-mae` option. We report this result as a resource-constrained reproduction and do not interpret it as a strict comparison with a pretrained-MAE configuration.

Another important observation is that zero-shot large vision-language models, including Qwen3-VL-4B-Instruct and DeepSeek-VL2-Tiny, perform much worse than task-specific multimodal detectors. Although these models possess strong general multimodal understanding ability, they do not establish a reliable decision boundary for fake news detection without task-specific adaptation. In contrast, GLED-Net benefits from explicit multimodal evidence modeling and supervised optimization.

Overall, the results establish strong predictive performance under the two Chinese benchmark settings. They do not by themselves isolate the effects of CLIP fine-tuning, the two streams, or adaptive fusion; these factors are evaluated through the controlled experiments below.

† The MAE checkpoint was unavailable in our environment, so the Bootstrap image encoder was randomly initialized using the `--no-pretrained-mae` option. This result is a resource-constrained reproduction and is not directly comparable with a pretrained-MAE setting.

Ablation Study. Table 3 separates backbone adaptation from the two streams, branch supervision, and fusion. On Weibo-16, removing the local stream reduces F1 from 93.97% to 90.24%, while removing the global stream reduces it to 89.49%. The two views are therefore complementary on this dataset. Fixed mean fusion obtains 93.84% F1, showing only a small benefit from the learned gate. Freezing CLIP and replacing it with a legacy encoder cause larger drops, confirming that pretrained representation and task-specific fine-tuning are major sources of performance.

FNDC exhibits a different pattern. Removing the global stream reduces F1 from 98.29% to 96.57%, whereas removing the local stream produces 98.30%, which is effectively tied with the full model.

Table 3: Component ablation on Weibo-16 and FNDC (%). All variants use the same split and training protocol. The full model uses $K = 4$ and $\lambda = 0.2$. The best result in each column is bold.

	Weibo-16				FNDC			
Variant	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Full GLED-Net	**94.06**	**95.41**	**92.58**	**93.97**	98.30	98.99	97.60	98.29
Frozen CLIP	89.13	89.44	88.73	89.08	95.70	94.72	96.80	95.75
w/o auxiliary supervision	90.83	92.00	89.43	90.70	98.00	**99.59**	96.40	97.97
w/o local cue stream	90.66	94.46	86.38	90.24	98.30	98.40	**98.20**	98.30
w/o global narrative stream	89.83	92.54	86.64	89.49	96.60	97.36	95.80	96.57
w/o adaptive fusion (mean)	93.93	95.32	92.40	93.84	**98.50**	98.99	98.00	**98.49**
Legacy non-CLIP model	76.51	80.56	69.87	74.84	95.10	94.30	96.00	95.14

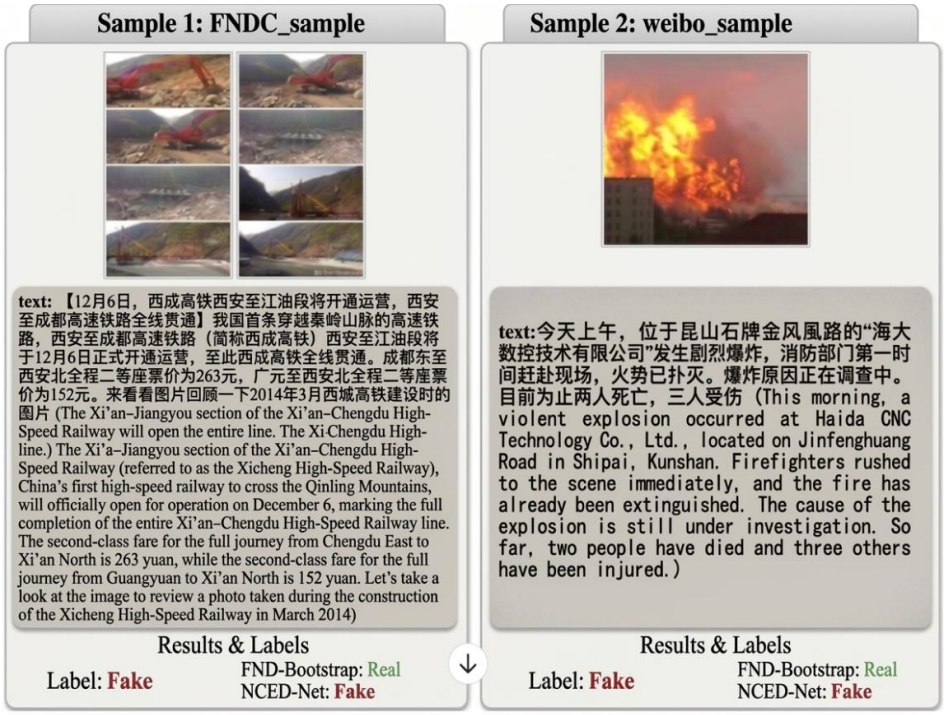

Figure 2: Illustrative FNDC cases. These examples provide a compact comparison between GLED-Net and FND-Bootstrap, but should not be interpreted as causal evidence or faithful explanations.

Fixed mean fusion reaches 98.49%, slightly above learned fusion. These single-run differences are small, but they show that the heuristic local stream and adaptive gate do not provide uniform gains across datasets. We therefore describe them as dataset-dependent design choices rather than universally beneficial mechanisms.

Qualitative Analysis on FNDC. We inspect both successful and failed predictions rather than selecting only cases where GLED-Net outperforms a baseline. Figure 3 visualizes the four FNDC examples summarized in the main text. Panels (a) and (b) are successful GLED-Net predictions for which FND-Bootstrap is incorrect. Panel (c) is a fake post for which both branches assign low fake probabilities. Panel (d) shows a branch-conflict failure: the global branch assigns a relatively high fake probability, whereas the local branch assigns a very low probability, and the fused prediction remains real. The reported gate value is the mean of the feature-wise global fusion weights. It is a diagnostic statistic rather than a calibrated measure of evidence reliability. Figure 2 provides the compact main-paper comparison, while Appendix Figure 3 presents the complete diagnostic panel, including the input images, ground-truth and model predictions, TF–IDF cue words, branch-level fake probabilities, final probability, and mean fusion weight. These examples are illustrative rather than causal evidence and do not establish that the learned gate is a faithful human-interpretable

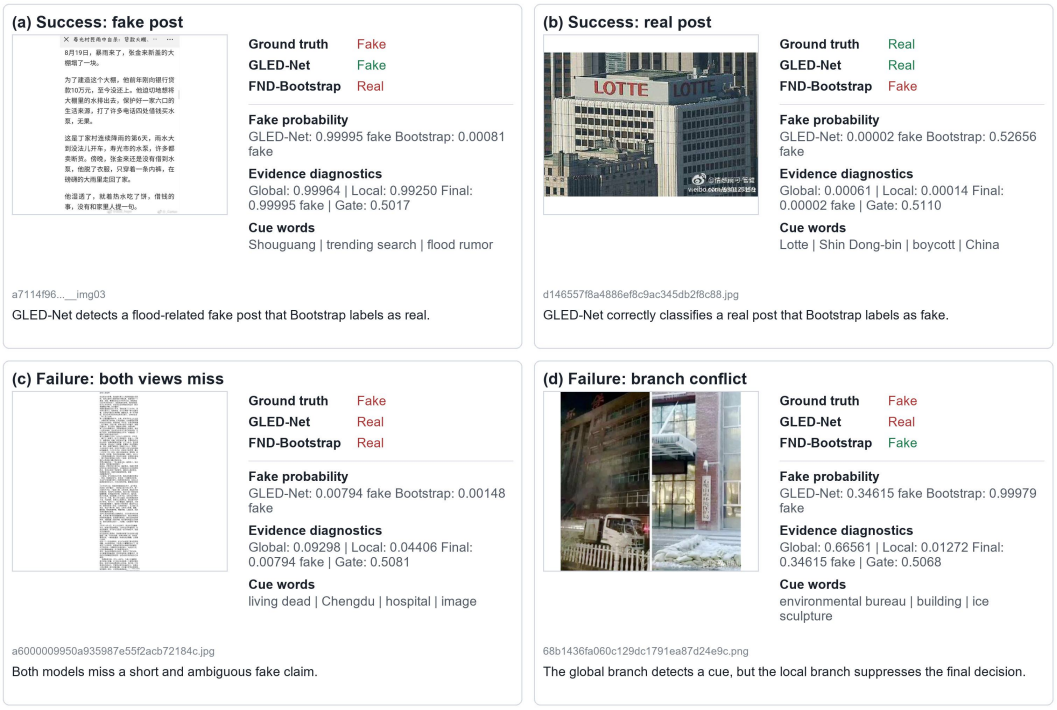

Figure 3: Qualitative FNDC examples containing two successful predictions and two failure cases. The panels report ground truth, GLED-Net and FND-Bootstrap predictions, branch and final fake probabilities, selected cue words, and the mean feature-wise fusion weight. These examples are illustrative and should not be interpreted as causal or faithful explanations.

reasoning process. The analysis is limited to FNDC and does not establish that the observed behavior generalizes to Weibo-16.

5 Conclusion

We presented GLED-Net, a CLIP-based framework that organizes multimodal fake news information into global narrative and local keyword-conditioned views. Experiments on Weibo-16 and FNDC show strong predictive performance, while the same-backbone and component controls clarify that CLIP adaptation is a major contributor. The local stream and learned fusion improve Weibo-16 but do not provide uniform gains on FNDC, where a global-only model is competitive and fixed mean fusion performs slightly better. These findings limit broad mechanistic claims and motivate adaptive, grounded, and human-verifiable cue selection, as well as evaluation on newer multilingual benchmarks.

AI use statement

Generative AI tools were used to assist with language editing and code debugging during manuscript preparation. The authors reviewed the resulting text, experimental code, and all reported claims, and take full responsibility for the final submission.

Ethics Statement

This work studies automatic multimodal misinformation detection and may support fact-checking or content-moderation workflows. It should not be used as a fully automated moderation decision system because false positives, dataset shift, and uneven performance across domains can affect legitimate content. We recommend human oversight, transparent deployment policies, and continuous

failure-case monitoring. The experiments use publicly available benchmark data and do not involve new human-subject data collection.

Reproducibility Statement

The manuscript specifies the model architecture, preprocessing procedure, hyperparameters, data splits, random seeds, and evaluation metrics. In particular, the FNDC construction procedure uses `piclist`-based image matching, class balancing, and a fixed seed, while the Weibo-16 seed study fixes the train/validation split. The controlled and component ablations use the same training protocol as the full model. These details are intended to support reproduction of the reported experiments.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 549–556, 2020. doi: 10.1609/aaai.v34i01.5393.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1351–1360, 2024. doi: 10.1145/3664647.3680663.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pp. 104–120, 2020. doi: 10.1007/978-3-030-58577-8_7.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pp. 2897–2905, 2022. doi: 10.1145/3485447.3511968.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, 2017a. doi: 10.1145/3123266.3123454.

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

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1760–1770, 2021. doi: 10.1109/ICCV48922.2021.00180.

KOTO. Fake news detection on chinese (zhiyuan & institute of computing technology, cas internet false news detection challenge dataset). `https://www.heywhale.com/mw/dataset/5e096a422823a10036b10a81`, 2019.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b. doi: 10.1109/CVPR52688.2022.01069.

Qilei Li, Mingliang Gao, Guisheng Zhang, Wenzhe Zhai, Jinyong Chen, and Gwanggil Jeon. Towards multimodal disinformation detection by vision-language knowledge interaction. *Information Fusion*, 102:102037, 2024. doi: 10.1016/j.inffus.2023.102037.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. FKA-Owl: Advancing multimodal fake news detection through knowledge-augmented LVLMs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10154–10163, 2024. doi: 10.1145/3664647.3681089.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM Web Conference 2025*, pp. 5139–5150, 2025. doi: 10.1145/3696410.3714522.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, 2021. doi: 10.18653/v1/2021.emnlp-main.545.

Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 708–717, 2017. doi: 10.18653/v1/P17-1066.

Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1980–1989, 2018. doi: 10.18653/v1/P18-1184.

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

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14444–14452, 2023. doi: 10.1609/AAAI.V37I12.26689.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024. doi: 10.1109/CVPR52733.2024.01240.

Jiao Qiao, Xianghua Li, Chao Gao, Lianwei Wu, Junwei Feng, and Zhen Wang. Improving multimodal fake news detection by leveraging cross-modal content correlation. *Information Processing & Management*, 62(5):104120, 2025. doi: 10.1016/j.ipm.2025.104120.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2023.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36, 2017. doi: 10.1145/3137597.3137600.

Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 312–320, 2019. doi: 10.1145/3289600.3290994.

Sahar Tahmasebi, Eric Muller-Budack, and Ralph Ewerth. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2189–2199, 2024. doi: 10.1145/3627673.3679826.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 5099–5110, 2019. doi: 10.18653/v1/D19-1514.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. MMDFND: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1178–1186, 2024. doi: 10.1145/3664647.3681317.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 849–857, 2018. doi: 10.1145/3219819.3219903.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 540–547, 2020. doi: 10.1145/3372278.3390713.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2560–2569, 2021. doi: 10.18653/v1/2021.findings-acl.226.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese CLIP: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. doi: 10.48550/arXiv.2211.01335.

Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 1417–1423, 2020. doi: 10.24963/ijcai.2020/197.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pp. 5384–5392, 2023.

Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10467–10484, 2024.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. MFAN: Multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 2413–2419, 2022. doi: 10.24963/ijcai.2022/335.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022. doi: 10.1109/CVPR52688.2022.01629.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pp. 2825–2830. IEEE, 2023.

APPENDIX

A Evaluation Metrics

Considering that fake news detection is a binary classification task (real/fake), we adopt four standard evaluation metrics to comprehensively assess model performance: Accuracy, Precision, Recall, and F1 score. Their definitions are as follows:

- **Accuracy**: the proportion of correctly classified samples among all samples, reflecting the overall classification performance of the model. It is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

  where TP (True Positive) denotes the number of fake news samples correctly classified as fake, TN (True Negative) denotes the number of real news samples correctly classified as real, FP (False Positive) denotes the number of real news samples incorrectly classified as fake, and FN (False Negative) denotes the number of fake news samples incorrectly classified as real.
- **Precision**: the proportion of correctly predicted fake news samples among all samples predicted as fake, reflecting the model's ability to avoid misclassifying real news as fake. It is defined as

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{9}$$

- **Recall**: the proportion of correctly predicted fake news samples among all actual fake news samples, reflecting the model's ability to detect fake news. It is defined as

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{10}$$

- **F1 score**: the harmonic mean of Precision and Recall, which balances the two metrics and provides a comprehensive evaluation of classification performance. It is defined as

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{11}$$

Since fake news detection requires both high precision and high recall, we use F1 score as the primary evaluation metric in our experiments.

B Baselines

We compare our model with the following baselines:

- **Qwen3-VL-4B-Instruct** Bai et al. (2025) is an instruction-tuned 4B dense vision-language model from the Qwen3-VL family. It supports joint understanding of image and text inputs with strong multimodal reasoning and long-context capabilities. In our experiments, we use Qwen3-VL-4B-Instruct as a zero-shot multimodal baseline: the model directly receives the news image and the corresponding post text, and is prompted to predict whether the sample is `real` or `fake` in a structured JSON format, without any task-specific fine-tuning.
- **DeepSeek-VL2-Tiny** Wu et al. (2024) is the lightweight variant of DeepSeek-VL2, a mixture-of-experts vision-language model designed for advanced multimodal understanding. It improves visual comprehension through dynamic tiling and enhances language-side efficiency with the DeepSeekMoE backbone. In our experiments, we use DeepSeek-VL2-Tiny as a zero-shot multimodal baseline under the same protocol as Qwen3-VL-4B-Instruct: the model takes the image and the associated post text as input and directly outputs a JSON prediction of `real` or `fake`, without further fine-tuning on the target dataset.
- **FND-Bootstrap** Ying et al. (2023) is a novel scheme of Bootstrapping Multi-view Representations (BMR) for fake news detection. Given a multi-modal news, FND-Bootstrap extracts representations respectively from the views of the text, the image pattern, and the image semantics. FND-Bootstrap introduces improved Multi-gate Mixture-of-Expert networks (iMMoE) for feature refinement and fusion. Representations from each view are separately used to coarsely predict the fidelity of the whole news, and the multimodal representations are able to predict the cross-modal consistency. With the prediction scores, FND-Bootstrap reweighs each view of the representations and bootstraps them for fake news detection.

Table 4: Random-seed stability on Weibo-16 (%). We report mean ± standard deviation over five runs with a fixed train/validation split.

Setting	Accuracy	Precision	Recall	F1
Mean ± Std	93.21 ± 0.62	95.03 ± 0.33	91.18 ± 1.26	93.06 ± 0.67

- **FND-CLIP** Zhou et al. (2023) is a multimodal Fake News Detection network based on Contrastive Language-Image Pretraining (CLIP). FND-CLIP extracts the deep representations together from news using two unimodal encoders and two pair-wise CLIP encoders. The CLIP-generated multimodal features are weighted by the CLIP similarity of the two modalities. FND-CLIP further introduces a modality-wise attention module to aggregate the features.
- **MIMoE-FND** Liu et al. (2025) improves the performance of multimodal fake news detection by integrating a mixture-of-experts (MoE) architecture with cross-modal contrastive learning. It leverages multiple specialized expert sub-networks to separately process different types of multimodal features (textual, visual, and auxiliary context features), and dynamically assigns input features to the most suitable experts based on their characteristics. This design enables the model to fully exploit the complementary information of different modalities while avoiding information redundancy and noise interference. By combining MoE-based feature specialization with contrastive learning, MIMoE-FND enhances the model's ability to distinguish real and fake content, effectively capturing subtle inconsistencies between text and image in complex fake news scenarios.

C Random-Seed Stability

To evaluate whether the performance of our method is robust to optimization randomness, we repeat the full training and testing pipeline five times on Weibo-16 using different random seeds $\{42, 3407, 2024, 2025, 7\}$. To avoid conflating optimization noise with data-split variation, we fix the train/validation split and change only the random seed used for model initialization, data loading, and training order. All other hyperparameters follow the default setting of the main experiments. Due to computational constraints, we report this analysis on Weibo-16 only.

Table 4 reports the mean and standard deviation over five runs. GLED-Net achieves $93.21\% \pm 0.62\%$ Accuracy and $93.06\% \pm 0.67\%$ F1. Precision has the smallest variation, whereas Recall varies more across runs. These values characterize optimization stability and do not constitute a significance test against baselines, which were not repeated under matched seeds.

D Hyperparameter Sensitivity

We analyze the auxiliary-loss weight λ on Weibo-16 and the keyword budget K on both datasets. We vary one hyperparameter at a time, keep the train/validation split fixed, and train each configuration for 5 epochs with seed 42.

Table 5 shows the effect of the auxiliary-loss weight. Among the tested settings, the best result is achieved at $\lambda = 0.2$, yielding 94.06% Accuracy and 93.97% F1. The settings $\lambda = 0.3$ and $\lambda = 0.4$ remain competitive, indicating that the model is not overly sensitive to the auxiliary-loss weight within a moderate range. In contrast, removing auxiliary supervision entirely ($\lambda = 0.0$) or assigning it too little weight leads to weaker performance, suggesting that proper auxiliary supervision is important for preserving branch discriminability.

Table 6 reports sensitivity to the keyword budget. $K = 4$ gives the highest F1 on Weibo-16, whereas $K = 8$ gives the highest F1 on FNDC. The differences among $K \in \{1, 4, 8\}$ are modest, particularly on FNDC. The result supports robustness to a reasonable keyword budget rather than proving that a specific number of terms constitutes faithful evidence. Together with the component ablation, it suggests that the local stream can behave as a compact complementary text representation and may not always isolate decisive cues.

Table 5: Sensitivity to the auxiliary-loss weight λ on Weibo-16 (%). We fix $K = 4$ and train each configuration for 5 epochs.

λ	Accuracy	Precision	Recall	F1
0.0	91.22	95.04	86.99	90.83
0.05	92.36	**96.28**	88.12	92.02
0.1	92.27	92.83	91.62	92.22
0.2	**94.06**	95.41	92.58	**93.97**
0.3	93.36	93.14	**93.62**	93.38
0.4	92.88	92.18	93.71	92.94

Table 6: Sensitivity to keyword budget K on Weibo-16 and FNDC (%). We fix $\lambda = 0.2$ and train each configuration for 5 epochs.

	Weibo-16				**FNDC**			
K	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
1	93.36	94.53	92.05	93.27	98.30	98.79	97.80	98.29
4	**94.06**	**95.41**	**92.58**	**93.97**	98.30	98.99	97.60	98.29
8	93.36	95.18	91.35	93.23	**98.60**	**99.19**	**98.00**	**98.59**

E Controlled Comparison with the Same Backbone

To verify that the performance gain of GLED-Net does not merely come from using a strong Chinese-CLIP backbone, we further construct a controlled baseline with the same backbone and training protocol. Specifically, we keep the same Chinese-CLIP encoder, data split, optimization settings, and training schedule unchanged, but remove the proposed evidence decomposition and adaptive gated fusion. Instead, the controlled baseline performs a simple single-branch fusion over the image, the raw text, the keyword representation, and their discrepancy features, followed by a standard classifier. In this way, the only essential difference between the two models is whether the proposed decomposition-and-fusion design is used.

Table 7 reports the Weibo-16 comparison. The controlled baseline obtains 93.19% Accuracy and 93.03% F1, confirming that Chinese CLIP is itself a major contributor. GLED-Net improves these values to 94.06% and 93.97%, respectively, under the same backbone. This supports an additional contribution from the complete organization on Weibo-16, but does not imply that every component is universally useful; the FNDC ablation in Table 3 shows that fixed mean fusion and the global-only variant are highly competitive.

F Backbone Generality

We further evaluate whether GLED-Net can generalize across different vision-language backbones. Specifically, we keep the overall architecture of GLED-Net unchanged and only replace the underlying multimodal encoder. In addition to the default Chinese-CLIP backbone, we instantiate GLED-Net with three alternative backbones: BLIP, SigLIP2, and AltCLIP. For a controlled comparison, all variants are trained and evaluated on the same Weibo-16 benchmark under the same data setting, and only the backbone is changed.

Table 8 reports the comparison results. Among the alternative backbones, AltCLIP achieves the best overall performance, reaching 91.05% Accuracy and 90.88% F1, while BLIP attains 88.08% Accuracy and 88.49% F1, and SigLIP2 yields 87.51% Accuracy and 86.72% F1. These results show that GLED-Net can be successfully instantiated on multiple vision-language encoders and remains effective beyond the default Chinese-CLIP setting, which supports the backbone-level generality of the proposed decomposition-and-fusion framework.

Table 7: Controlled comparison with the same Chinese-CLIP backbone on Weibo-16 (%). The controlled baseline removes the proposed evidence decomposition and adaptive gated fusion, and replaces them with a simple single-branch fusion module.

Method	Accuracy	Precision	Recall	F1
CLIP-Control	93.19	95.24	90.92	93.03
GLED-Net	**94.06**	**95.41**	**92.58**	**93.97**

Table 8: Backbone comparison on Weibo-16 (%). We keep the overall GLED-Net architecture unchanged and replace only the multimodal backbone.

Backbone	Accuracy	Precision	Recall	F1
BLIP	88.08	85.56	91.62	88.49
SigLIP2	87.51	92.57	81.57	86.72
AltCLIP	91.05	92.65	89.17	90.88
Chinese-CLIP (default)	**94.06**	**95.41**	**92.58**	**93.97**

At the same time, the default Chinese-CLIP backbone still delivers the strongest performance, substantially outperforming all three alternatives. We attribute this to two reasons. First, Chinese-CLIP is better aligned with the linguistic characteristics of Chinese social-media posts, which makes it more suitable for Weibo-16 than general-purpose multilingual or English-centric models. Second, the proposed framework benefits from a backbone whose image-text embedding space is already well aligned for Chinese short-text understanding. Nevertheless, the fact that GLED-Net remains competitive across BLIP, SigLIP2, and AltCLIP indicates that the effectiveness of our method is not tied to a single backbone implementation.

G LIMITATIONS

GLED-Net has several limitations. First, the evaluation uses only two Chinese datasets, and both are relatively old. We therefore do not establish cross-language, cross-platform, recent-event, or out-of-context generalization. Event-Radar, ViKI, HAMMER, SNIFFER, NewsCLIPpings, and VERITE were not evaluated under the same protocol, so no superiority claim is made against those methods or settings.

Second, the local stream relies on TF–IDF keywords extracted by Jieba. These terms are not human-annotated evidence and may omit decisive entities or retain spurious words. Its weak sensitivity to K and the FNDC global-only result suggest that it can sometimes act as a redundant compact text encoding rather than a precise evidence selector. Future work should evaluate cue quality against human annotations and explore adaptive or grounded selection.

Third, the FNDC results use the fixed sample-level split produced after piclist-based image matching and multi-image expansion. We do not perform an additional grouped-by-original-post leakage audit in this revision; duplicated source text or images, if present in the released records, could make the reported estimate optimistic. The FNDC numbers should therefore be interpreted under the stated preprocessing and split protocol.

Fourth, feature gating, discrepancy concatenation, auxiliary supervision, adaptive fusion, and TF–IDF keyword extraction are established components. Our contribution is their task-specific organization into global and local views, not the novelty of the individual operators. Keywords are heuristic and not human-verified evidence; the local branch may behave like a compact text representation, and fusion behavior is dataset-dependent. The FNDC fixed- fusion result also shows that adaptive fusion is not uniformly beneficial.

Finally, the model handles a single image–text pair and does not use external knowledge, temporal propagation, user context, or social structure. We do not evaluate calibration, fairness, adversarial robustness, or distribution shift. The case studies are illustrative rather than causal explanations. The FND-Bootstrap reproduction also used a randomly initialized MAE because the pretrained checkpoint was unavailable.

H Broader Impact

This work has potential positive societal impact in that it may help improve the automatic detection of multimodal misinformation on social-media platforms. More reliable fake news detection systems could assist content moderation, support fact-checking workflows, reduce the spread of harmful misinformation, and improve the quality of online information ecosystems. In particular, methods that better capture subtle cross-modal inconsistencies may be useful in scenarios where misleading content is difficult to detect through text-only or image-only analysis.

At the same time, this work may also have negative societal implications if deployed without sufficient care. First, like other automated fake news detection systems, our method may produce false positives, incorrectly flagging legitimate content as fake, which could affect freedom of expression, public trust, or fair content moderation. Second, model performance may vary across domains, events, platforms, or linguistic styles, which raises the risk of distribution shift and uneven performance across different user groups or content sources. Third, such systems could be misused in overly aggressive or opaque moderation pipelines without appropriate human oversight and appeal mechanisms. Finally, although our work focuses on misinformation detection rather than content generation, stronger multimodal inconsistency modeling may also be studied or adapted in adversarial settings by malicious actors seeking to better understand the weaknesses of current detection systems.

For these reasons, we believe that practical deployment of multimodal fake news detection systems should include careful human oversight, transparent usage policies, domain-specific evaluation, and continuous monitoring for failure cases and unintended bias. We view this work as a step toward more reliable multimodal misinformation analysis, but not as a complete solution for high-stakes automated moderation.

I Compute Resources

All experiments were conducted on a workstation equipped with one NVIDIA RTX 4090D GPU and one NVIDIA RTX 3090 GPU, with 48 GB aggregate GPU memory. A single main GLED-Net run required approximately 6 hours per dataset, while a single ablation run required approximately 1 hour. These are single-run wall-clock measurements under the corresponding settings.

The random-seed stability, hyperparameter sensitivity, controlled same-backbone comparison, and backbone generality experiments were conducted under the same hardware setting. Since the paper includes multiple main experiments, ablation experiments, and supplementary analyses, the total compute cost of the full study is substantially larger than that of a single reported run. These details are provided to facilitate a more realistic understanding of the computational requirements for reproducing the reported results.