Falsdo: Benchmarking Artifact-Controlled Multimodal Fake News Verification via Failure-Aligned Auditing

Bowen Chen[1], **Jun Yin**[2], **Lele Cao**[3], **Zheyuan Zhan**[1] and **Can Wang**[1*]
[1]Zhejiang University
[2]Tsinghua University
[3]Beijing Normal University
bwchen@zju.edu.cn, yinj24@mails.tsinghua.edu.cn, caolele_happy@mail.bnu.edu.cn,
632468048@qq.com, wcan@zju.edu.cn

Abstract

Recent generative AI renders multimodal misinformation structurally harder to detect, making reliable detection dependent on semantic verification grounded in verifiable evidence. However, current benchmarks often fail to isolate true semantic checking from superficial shortcut exploitation. We introduce FALSDO, a diagnostic benchmark designed to make robustness and auditability identifiable. Constructed via a heterogeneous web-mining pipeline, FALSDO provides instruction-conditioned grounding and formalizes a counterfactual artifact-control protocol. This stress test reveals a critical failure: when low-level generation traces are equalized, detector performance degrades, indicating reliance on artifacts rather than robust verification. To enable reproducible diagnosis, we propose **DREA**, a failure-aligned evidence-auditing baseline. DREA specifies evidence acquisition with explicit noise control and implements constrained channel-wise auditing with reliability-aware late fusion. Experiments demonstrate that FALSDO reliably exposes shortcut dependence, while DREA improves instruction-level grounding (Joint-F1) and minimizes degradation under artifact-controlled settings.

1 Introduction

In recent years, fake news has increasingly evolved into complex forms, combining textual narratives with highly realistic images or AI-generated content. Such multimodal content poses a severe threat, as visual cues significantly enhance perceived credibility [Newman and Schwarz, 2024; Pulm *et al.*, 2024]. Moreover, emotional rendering in these fabrications further exacerbates users' gullibility and accelerates the dissemination of deceptive content [Ali Adeeb and Mirhoseini, 2023; Vosoughi *et al.*, 2018]. With the rapid growth of synthetic media in social ecosystems [Bui *et al.*, 2024; Wilson *et al.*, 2023; Mirsky and Lee, 2021], traditional single-modal detection or shallow fusion methods struggle to effectively cope with the structural complexity of multimodal fake news in real-world scenarios.

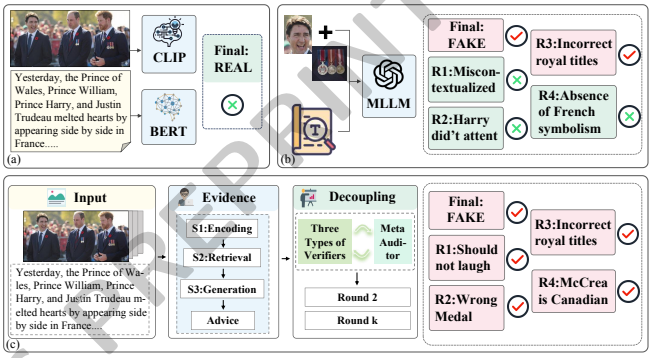


Figure 1: Comparison of multimodal fake news detection paradigms. (a) Traditional methods: simple feature fusion. (b) General MLLMs: one-pass parametric reasoning. (c) **Reference evidence auditing baseline**: constrained channel-wise checks with reliability-aware fusion (instantiated by DREA in this paper).

To address this, researchers have proposed various multimodal detection methods, typically attempting end-to-end discrimination through feature fusion. However, as illustrated in Figure 1, existing paradigms face limitations when heterogeneous tasks are compressed into the same discrimination channel. In high-stakes scenarios requiring the synthesis of multi-source evidence, these models often fail to capture fine-grained manipulation clues stably, lacking both robustness and interpretability.

These issues are compounded by insufficient characterization at the evaluation level. While existing datasets have laid a foundation [Boididou *et al.*, 2018; Jin *et al.*, 2017; Wang, 2017], most focus on short text or limited social media contexts. Many benchmarks merely require models to judge image-text consistency rather than factual veracity [Shu *et al.*, 2020]. Even recent works introducing external knowledge [Luo *et al.*, 2021; Shao *et al.*, 2023; Papadopoulos *et al.*, 2024; Schlichtkrull *et al.*, 2023] are dominated by overall authenticity judgments [Thorne *et al.*, 2018; Augenstein *et al.*, 2019]. Consequently, a critical *evaluation confound* remains: it is difficult to distinguish whether a model truly performs evidence grounding or merely exploits statistical

*Corresponding author.

Figure 2: Falsdo dataset example: Showing fine-grained visual manipulations (e.g., temporal displacement, unauthorized medal wearing) and textual manipulations (e.g., factual errors, fabricated details) in the dataset.

shortcuts (fake patterns) in the data [Geirhos *et al.*, 2020; Arjovsky *et al.*, 2020].

To bridge this gap, we construct FALSDO (**F**ine-grained **A**nalysis of **L**ogical and **S**emantic **D**eception in **O**nline content), a diagnostic benchmark designed to systematically measure robust, grounded semantic verification. As shown in Figure 2, FALSDO differs from binary-label datasets by providing instruction-based fine-grained annotations. Importantly, each instruction is paired with explicit grounding evidence (image boxes / text spans), aligning our benchmark with established region-level phrase grounding settings [Yu *et al.*, 2016; Mao *et al.*, 2016]. Crucially, we introduce an Artifact Control protocol: a counterfactual stress test that equalizes low-level generation traces while preserving event semantics. Under this setting, we observe that monolithic multimodal validators tend to over-rely on shortcut cues learned from generation artifacts, and even unconstrained verification pipelines can drift toward shortcut-dominated consensus under noisy evidence.

These observations motivate a *failure-aligned* design: semantic verification should be decomposed into representation-incompatible channels. We specify **DREA** (**D**ecoupled **R**easoning **E**xperts with **A**daptive Arbitration), a reference evidence-auditing system (Figure 1c). DREA treats detection as an investigation rather than static classification. It processes textual logic checking, visual forensic analysis, and cross-modal consistency through independent reasoning channels, all guided by retrieval-augmented evidence. Finally, an adaptive fusion mechanism arbitrates these multi-perspective conclusions to output an explained verdict. DREA serves as a reproducible baseline for instantiating the benchmark, not as a standalone novelty claim.

The main contributions of this paper are summarized as follows:

- We implement a web-mining pipeline with instruction-level grounding, advancing evaluation from label accuracy to verifiable evidence.
- We construct FALSDO as a diagnostic benchmark with a logic-centric taxonomy, long-form narratives, and counterfactual artifact-control evaluation that makes semantic robustness directly measurable.
- We provide DREA as a *reference evidence-auditing baseline* that specifies evidence acquisition and constrained channel-wise auditing with reliability-aware fusion.
- FALSDO exposes shortcut collapse under artifact equalization; strong auditing baselines improve grounding and reduce degradation in the artifact-controlled setting.

2 Related Work

Robust and Grounded Multimodal Verification. High accuracy on unconstrained benchmarks often masks brittle behavior, where models rely on spurious correlations rather than true understanding [Geirhos *et al.*, 2020; Sagawa *et al.*, 2020; Arjovsky *et al.*, 2020]. For multimodal misinformation, auditability requires grounded verification, localizing specific image regions and text spans [Shao *et al.*, 2023; Yu *et al.*, 2016; Krishna *et al.*, 2017], instead of producing single holistic labels. However, most fake news setups emphasize end-to-end veracity, obscuring the distinction between genuine semantic checking and shortcut exploitation [Shu *et al.*, 2020]. Our FALSDO benchmark bridges this gap by pairing instruction-level grounding with a counterfactual artifact-control protocol to directly test semantic robustness.

Multimodal Fake News Detection. Early detection methods relied on feature fusion for overall discrimination, such as EANN [Wang *et al.*, 2018], MVAE [Khattar *et al.*, 2019], MGN [Qian *et al.*, 2021], SpotFake [Singhal *et al.*, 2019], and Fakeddit [Nakamura *et al.*, 2020]. Later works modeled fine-grained associations through cross-modal attention [Wu *et al.*, 2021; Chen *et al.*, 2022]. Recent research attempts to introduce localization; for instance, HAMMER [Shao *et al.*, 2023] grounds tokens and CMA-VRP [Li *et al.*, 2025] reveals inconsistencies. However, these methods are still dominated by overall judgments, making it difficult to systematically characterize complex, multi-location semantic manipulations.

Retrieval-Augmented and Agentic Verification. RAG [Lewis *et al.*, 2020; Karpukhin *et al.*, 2020] introduces external knowledge, but verification is often bottlenecked by evidence noise. While RARG [Yue *et al.*, 2024], DEFAME [Braun *et al.*, 2024], and vision-native indexing [Yu *et al.*, 2024] extend this to multimodal scenarios, distinguishing robust reasoning from artifact shortcuts remains challenging. Concurrently, multi-agent frameworks like MIRAGE [Shopnil *et al.*, 2025], RAMA [Yang *et al.*, 2025], EXCLAIM [Wu *et al.*, 2025], and MAD-Sherlock [Lakara *et al.*, 2025] focus on generating interpretable verdicts. We complement them by emphasizing diagnostic evaluation. Our proposed DREA serves as a principled reference system, reducing shortcut-driven convergence via *artifact-only* information propagation and reliability routing.

3 FALSDO Dataset

We propose FALSDO, a multimodal fake news evaluation benchmark oriented towards *semantic and logical inconsistencies* and robust, grounded verification under counterfactual artifact equalization. FALSDO contains **12,740** high-quality image-text pairs, equipped with long-form narratives and dense instruction-conditioned annotations.

Unlike datasets primarily relying on visual artifact clues or coarse-grained image-text mismatches, FALSDO jointly provides in a unified benchmark: (i) a fine-grained manipulation taxonomy oriented towards world knowledge and narrative reasoning; (ii) instruction-level grounding evidence (visual box/text span) corresponding to each manipulation instruction; (iii) an artifact-control comparative setting to suppress low-level generation trace shortcuts, thereby measuring semantic robustness and evidence grounding in a way that is directly auditable.

3.1 Systematic Data Acquisition Pipeline

To ensure high semantic diversity, we implement a staged web-mining pipeline targeting three heterogeneous sources: Authoritative News, including Reuters, The Washington Post, The Guardian, and AP News (60%); Collaborative Knowledge Bases such as Wikipedia (20%); and Professional Fact-checking Repositories like Snopes and FactCheck.org (20%).

We employ on-demand semantic load balancing to prevent overfitting to hot topics. Specifically, we enforce a coarse domain distribution during crawling: Protocol (40%), Spatio-Temporal (30%), and Context (30%). To operationalize this distribution, we constructed taxonomy-anchored seed lexicons for each domain. Specifically, we curated a set of high-frequency discriminative queries like bilateral treaty for Protocol, historical landmark for Spatio-Temporal, to perform targeted query-driven retrieval. This ensures that the retrieved raw corpus is explicitly aligned with our predefined semantic typology, rather than relying on random web crawling. To support complex reasoning, only samples with captions exceeding **100 words** are retained. Crucially, we maintain a pure real distribution via dual near-duplicate de-duplication: (i) *Image-level* using Perceptual Hashing ($d_{img} = 6$), and (ii) *Text-level* using semantic similarity filtering ($\tau_{text} = 0.92$). This guarantees that the base samples are distinct and high-quality.

3.2 Semantics-Driven Manipulation Generation

FALSDO adopts a semantics-driven plan-and-execute paradigm, emphasizing high semantic impact and low perceptual distortion. The generation process follows a strict protocol to simulate sophisticated fabrication tactics:

1. **Intent-conditioned Planning**: A SOTA Large Language Model acts as a planner to output a multi-step manipulation plan. We enforce a complexity constraint: each sample must contain **3** to **6** modifications, including at least one image edit and one text edit. The planner prioritizes *logical contradictions* ,e.g., protocol/fact/causality violations, over random noise.
2. **Instruction-based Multimodal Execution**: The visual side employs text-guided inpainting/completion [Ho *et al.*, 2020]; the text side executes counterfactual rewriting to flip key assertions while maintaining narrative fluency. This creates concealed but critical cross-modal conflicts.
3. **Instruction-level Grounding & Sanitization**: Each manipulation step generates explicit evidence: visual grounding boxes or text spans. We apply geometric sanitization, e.g., validity checks $y_{max} > y_{min}$, to ensure all grounding annotations are physically valid and auditable.

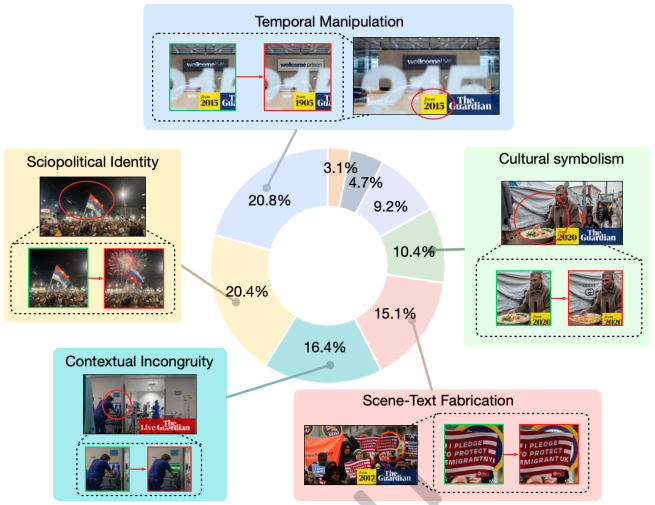


Figure 3: Category distribution of visual manipulations in FALSDO.

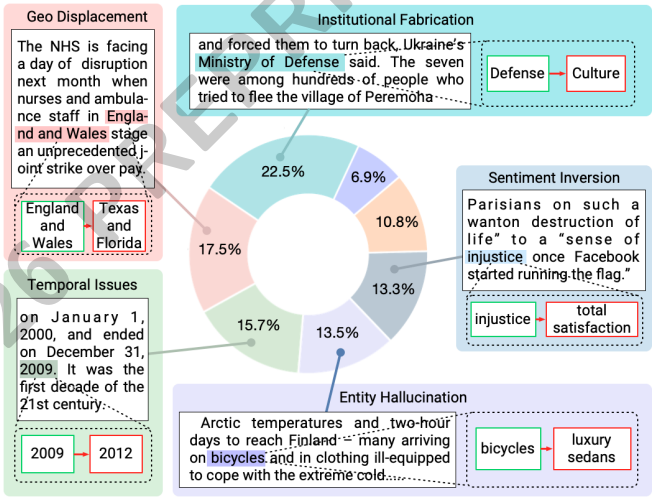
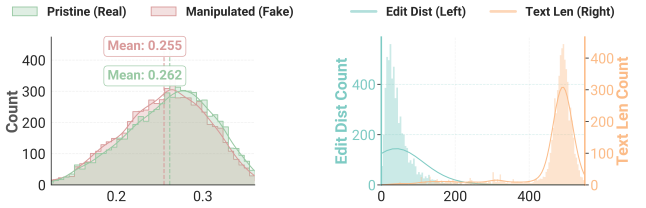


Figure 4: Category distribution of textual manipulations in FALSDO.

To strictly isolate semantic inconsistencies from the artifacts introduced during the generation process, we generate a *Control Group* by applying the same inpainting backend to Real images using an *identity-preserving prompt*, followed by identical post-processing, e.g., JPEG compression $Q = 85$. This protocol ensures that low-level generation traces are equalized across Real (Control) and Fake samples. This forces the model to ignore pixel-level artifacts and focus

Dataset	Source	Modality	Gen. Method	Granularity	Volume	Ground	Art. Ctrl	Long	Tax.
LIAR	PolitiFact	T	Fact-Check	Multi-class	12.8k	No	No	No	No
Weibo	Social Media	M	Crowd	Binary	9.5k	No	No	No	No
Twitter	Social Media	M	Crowd	Binary	15.6k	No	No	No	No
FakeNewsNet	Fact-Check	M	Editor-Verified	Binary	23k	No	No	No	No
MCFEND	Social Media	M	Agency-Verified	Binary	23k	No	No	No	No
DGM[4]	News	M	GAN	Multi-class	230k	Part.	No	Limited	Limited
NewsCLIPpings	News	M	Retrieval	Binary	988k	No	No	No	No
FALSDO (Ours)	**News/Mixed**	**M**	**LLM+Diffusion**	**Fine-Grained**	**12.7k**	**Yes**	**Yes**	**Yes**	**Yes**

Table 1: Comparison of FALSDO with existing fake news datasets. **Modality**: T (Text), I (Image), M (Multimodal). **Ground**: instruction-level grounding evaluation; **Art. Ctrl**: artifact-control setting; **Long**: long-form narrative; **Tax.**: manipulation taxonomy. **Part.** indicates partial support (e.g., grounding without instruction queries).



(a) CLIP Score Comparison (b) Text Length vs. Edit Distance

Figure 5: Dataset statistical characteristics: (a) The CLIP similarity distribution of original and manipulated images overlaps highly, indicating manipulations preserve global alignment; (b) The scatter relationship between text length and edit distance reflects the concealed nature where "minor changes trigger major semantic reversals".

on high-level semantic logic inconsistencies. Thus, FALSDO explicitly measures *reasoning ability under Artifact Control*, rather than simple knowledge memorization or statistical pattern matching. The protocol is a standardized diagnostic stress test for one controlled generation pipeline, not a guarantee over all editing backends.

3.3 Fine-Grained Taxonomy

To simulate the complexity of real-world multimodal misinformation, where fabrications often retain high semantic consistency while violating specific world knowledge, FALSDO introduces a comprehensive taxonomy of **15** fine-grained categories. As shown in Figure 3 and Figure 4, these categories encompass both visual and textual modalities, shifting the detection focus from low-level artifacts to reasoning over cross-modal discrepancies and narrative logic.

Visual Manipulation Dimensions. Visual modifications fundamentally restructure the narrative context by introducing semantic conflicts into the scene. **Temporal & Chronological Manipulation** (21%) stands out as the most prevalent category, targeting the scene's temporal integrity by falsifying textual date markers or introducing anachronisms that contradict the event's historical grounding. A comparable proportion is attributed to **Sociopolitical Identity Tampering** (20%), which recontextualizes the event's ideological stance through the precise modification of cultural symbols and protest signage. Additionally, **Contextual Object Incongruity** (16%) presents a unique challenge to common sense reasoning by embedding objects that disrupt the scene's established atmospheric or physical causality.

Textual Manipulation Dimensions. Textual fabrications aim to distort the precise factual constraints and named entities within the narrative. We observe that **Institutional & Protocol Fabrication** (23%) constitutes the primary textual distortion, involving the hallucination of incompatible administrative roles or entities to violate established diplomatic protocols. Ranking second, **Geospatial & Geopolitical Displacement** (18%) generates location-based conflicts by attributing events to geographically or geopolitically incongruent regions. Finally, a significant volume of cases involves **Chronological & Temporal Inconsistency** (16%), which disrupts the narrative timeline by asserting dates or durations that are factually irreconcilable with the visual evidence.

3.4 Dataset Analysis and Characterization

Dataset Statistics. FALSDO ultimately contains **12,740** class-balanced samples (Train: 10,192 / Val: 1,274 / Test: 1,274). Figure 5(a) shows that the CLIP similarity distributions [Radford *et al.*, 2021] of original and manipulated samples overlap highly, indicating that manipulations preserve global alignment. Figure 5(b) shows that under long texts (Avg. 465 words), small edit distances can trigger significant factual reversals, reflecting the core difficulty of the benchmark.

Comparison with Existing Datasets. Table 1 compares FALSDO with mainstream benchmarks. LIAR, FakeNewsNet, etc., are mostly coarse-grained binary classifications; NewsCLIPpings focuses on out-of-context pairing. In contrast, FALSDO is a benchmark that jointly supports instruction-level grounding, artifact-control evaluation, long-form narratives, and a logic-centric taxonomy, systematically measuring the robustness of semantic verification.

Quality Control and Auditability. We use 500 aligned Real/Control/Fake triplets for QA. Manual validation on 100 triplets yields 99% semantic preservation, 100% contradiction validity, and 99% grounding reasonableness. Sanitization discards 6.2% of samples; the geometric-sanitization false-positive rate is 2.8%, the control semantic failure rate is 4.5%, and accuracy varies by at most 0.7 points across JPEG

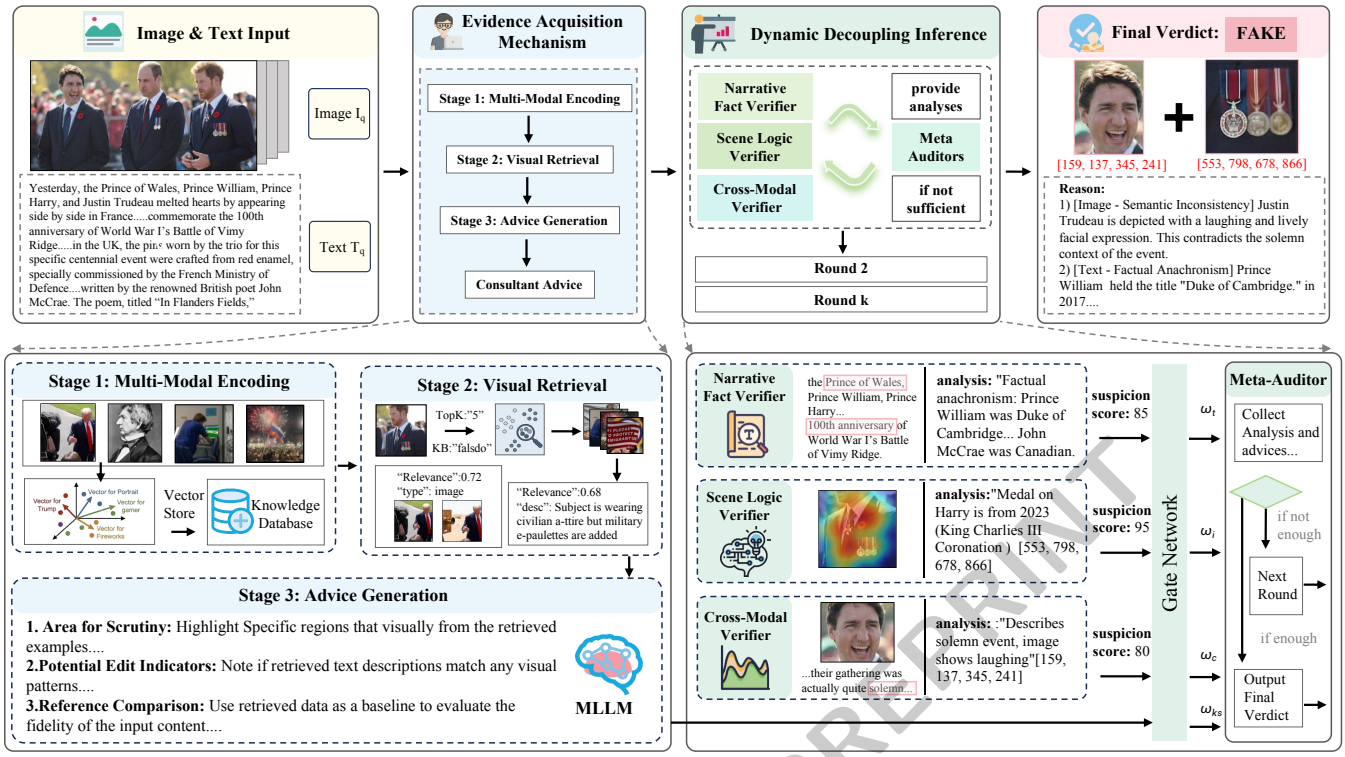


Figure 6: The architecture of DREA. (1) **Evidence Acquisition Mechanism (EAM)**: Retrieves multimodal evidence and synthesizes forensic advice with a kill-shot signal. (2) **Decoupled Verification**: Three independent verifiers scrutinize *narrative facts*, *cross-modal alignment*, and *scene logic* (focusing on physical/temporal affordances rather than artifacts). (3) **Reliability-Aware Arbitration**: A gating network dynamically weights (ω) the expert verdicts based on confidence to yield the final decision.

$Q = 70/85/95$.

4 Reference Auditing Baseline: DREA Specification

To instantiate FALSDO's diagnostic protocols with a reproducible baseline, we specify **DREA** (**D**ecoupled **R**easoning **E**xperts with **A**daptive Arbitration). We design DREA as a streamlined evidence-auditing specification. Unlike monolithic architectures, it focuses on structural mitigation of the failure modes observed in monolithic validators. As shown in Figure 6, DREA replaces end-to-end discrimination with a decoupled evidence-auditing pipeline defined by strict information flow constraints.

Diagnostic Target (Failure Modes). We explicitly design DREA to address five failures identified in FALSDO: **(F1) Representation Interference** in shared parameters; **(F2) Artifact Shortcut Learning** exposed by artifact-control; **(F3) Retrieval Noise**; **(F4) Evidence Conflict** across heterogeneous channels; and **(F5) Evidence Scarcity** in long-tail samples.

4.1 Problem Formulation

We model multimodal fake news detection as an evidence-driven binary classification task. Given a news sample $N = \{T, I\}$, where T is text and I is image, the system aims to predict the veracity label $y \in \{0, 1\}$ (0: Real, 1: Fake) and generate an explanation E grounded in verifiable evidence. Formally, the inference process is defined as a mapping:

$$E, y = \mathcal{M}_{meta}\big(\{O_k\}_{k=1}^{K}, \mathcal{F}_{eam}, \mathcal{A}_{eam} \mid N\big), \qquad (1)$$

where $\mathcal{M}_{meta}$ denotes the Meta-Auditor, $\{O_k\}$ represents the outputs from K decoupled expert channels, and $\{\mathcal{F}_{eam}, \mathcal{A}_{eam}\}$ represent structured retrieval artifacts.

4.2 Phase 1: Auditable Evidence Acquisition

Standard RAG often injects noise directly into reasoning. To mitigate **(F3) Retrieval Noise**, DREA enforces a *"Broad Entry, Strict Exit"* protocol.

Vision-Native Retrieval and Filtering. We encode the query q_N (using both image and text) and retrieve top-K candidates from the knowledge base $\mathcal{K}$ [Karpukhin *et al.*, 2020]. To prevent irrelevant context injection, we apply a strict exit threshold τ_{ret}:

$$\mathcal{C}_{filtered} = \{c_i \in \mathcal{K} \mid \text{sim}(q_N, c_i) > \tau_{ret}\}. \qquad (2)$$

If $\mathcal{C}_{filtered} = \emptyset$, the system executes a *Zero-Day Fallback*, blocking retrieval inputs to downstream experts to prevent hallucination.

The Evidence Acquisition Mechanism and Kill-Shot Scoring. Instead of end-to-end generation, an Evidence Acquisition Mechanism (EAM) distills $\mathcal{C}_{filtered}$ into Textual Facts ($\mathcal{F}_{eam}$) and a calibrated Kill-Shot Signal (s_{ks}). The Kill-Shot

signal quantifies the presence of decisive debunking patterns (e.g., known fake templates). We formulate s_{ks} as a continuous variable derived from retrieval similarity r_i and the EAM's pattern-match confidence $k_i \in [0, 1]$:

$$s_{ks} = \max_{c_i \in \mathcal{C}_{filtered}} (r_i \cdot k_i), \quad s_{ks} \leftarrow 0 \text{ if } \mathcal{C}_{filtered} = \emptyset. \quad (3)$$

This formulation turns decisive evidence into a differentiable signal that can be weighed against other channels, avoiding hard-coded rules.

4.3 Phase 2: Decoupled Auditing and Arbitration

To prevent **(F1) Representation Interference** and **(F2) Artifact Shortcut Learning**, DREA enforces deep decoupling: reasoning is split into representation-incompatible channels that communicate only via compact artifacts.

Independent Expert Channels. We instantiate three experts: Narrative Fact Verifier ($\mathcal{E}_{narr}$), Scene Logic Verifier ($\mathcal{E}_{scene}$), and Cross-Modal Verifier ($\mathcal{E}_{cm}$). Crucially, to ensure independence, experts receive verifiable facts $\mathcal{F}_{eam}$ but are blinded to the forensic advice $\mathcal{A}_{eam}$. Each expert k outputs a normalized suspicion score $\tilde{S}_k \in [0, 1]$ and a self-confidence score c_k.

Formalizing Reliability Arbitration. To handle **(F4) Evidence Conflict**, the final decision is mediated by a Meta-Auditor via a learned gating mechanism. Unlike static fusion, we propose a Sample-Adaptive Reliability Router g_ϕ that dynamically estimates the trustworthiness of each channel. We construct a routing feature vector $\mathbf{x}$ by concatenating expert confidences, embeddings, and the kill-shot score. The channel weights $\mathbf{w}$ are computed via a temperature-scaled Softmax:

$$\mathbf{w} = \text{softmax}\big(g_\phi(\mathbf{x})/\tau\big), \quad \mathbf{w} \in \mathbb{R}^{K+1}, \quad (4)$$

where τ controls the sharpness of the selection. The final verdict logic z is a reliability-weighted sum of expert scores and the kill-shot signal:

$$z = \sum_{k \in \text{Experts}} w_k \tilde{S}_k + w_{ks} s_{ks}. \quad (5)$$

Finally, the probability is given by $p_{meta} = \sigma(z)$. This mechanism ensures that if an image expert relies on artifacts (low reliability under control settings) or retrieval is noisy (low s_{ks}), the router can dynamically suppress their influence, achieving robustness.

Audit-Triggered Refinement. For **(F5) Evidence Scarcity**, we define a conditional feedback loop. If the Meta-Auditor's fusion confidence $c_{meta} = 2|p_{meta} - 0.5|$ falls below a safety threshold τ_{audit}:

$$\text{If } c_{meta} < \tau_{audit} \implies \text{Trigger}(\text{Refine} \rightarrow \mathcal{E}_{target}). \quad (6)$$

This triggers targeted natural language feedback to the most uncertain expert, simulating a "second look" mechanism to improve coverage on long-tail samples.

Model	Acc	Acc (C)	ΔAcc	Prec	Rec	F1	AUC
HAMMER	53.88	48.91	-4.97	54.37	47.52	50.71	52.89
FKA-Owl	55.10	55.05	-0.05	55.00	**89.80**	68.22	52.17
MIRAGE	59.20	54.34	-4.86	59.03	53.46	56.11	58.67
GPT-4o	67.34	65.50	-1.84	71.17	59.40	64.75	67.42
DEFAME	70.63	66.20	-4.43	66.29	86.43	**75.03**	65.61
RAMA	72.03	64.29	-7.74	72.14	63.87	67.76	71.53
Qwen3-VL	67.17	58.45	-8.72	68.09	64.78	66.39	67.18
Qwen3-VL+CoT-SC@5	70.50	61.80	-8.70	70.12	71.25	70.68	70.45
DREA (Ours)	**75.88**	**73.20**	**-2.68**	**75.31**	68.54	71.76	**75.18**

Table 2: Performance on Falsdo (%). C: artifact-controlled. ΔAcc = Acc(C) − Acc.

5 Experiment

To validate FALSDO as a diagnostic benchmark, we evaluate three dimensions: veracity (Acc/AUC), auditability (instruction-level grounding), and robustness (counterfactual artifact-control). Figure 7 visualizes the results.

Setup Protocol: All trainable baselines are supervised using the same SSP JSON targets and identical training budgets. We first apply SFT/LoRA [Hu *et al.*, 2021] to the backbone under SSP to obtain expert checkpoints, then train the reliability router (g_ϕ) with experts fixed. This ensures gains stem from *structural decoupling* rather than extra parameters.

5.1 Main Results: Diagnostics on Falsdo

Mitigating Shortcut Learning. As shown in Table 2, DREA achieves the highest Acc of 75.88% and AUC of 75.18% on Falsdo. Crucially, under the Artifact-Control (C) stress test, monolithic baselines FT-Qwen3-VL-8B (referred to as Qwen3-VL for brevity) and CoT-SC suffer a collapse of approximately **-8.7%**, revealing their heavy reliance on superficial generation traces. In contrast, DREA degrades by only **2.68%**. This diagnostic gap confirms that FALSDO's artifact-control protocol successfully exposes shortcut learning, and DREA's decoupled auditing effectively shifts reliance towards semantic consistency.

Budget-matched Efficiency. Compared to CoT-SC@5 which also uses 5 inference calls, DREA leads significantly with 75.88 vs. 70.50 while using fewer tokens and lower latency (4.5k tokens / 1.8s vs. 5.0k / 2.0s). This indicates that performance gains derive from evidence-guided structural auditing rather than simply scaling inference compute via multiple sampling.

The "Artifact Probe" Baseline. We treat the standalone Scene Logic Verifier ($\mathcal{E}_{scene}$) as an "artifact probe". It performs poorly on the original set scoring only 62.15 and deteriorates further under Control. This confirms that FALSDO cannot be solved by visual forensics alone, validating the dataset's emphasis on *semantic* rather than *pixel-level* manipulation.

5.2 Generalization to Public Benchmarks

DREA achieves optimal Acc on all four benchmarks (Table 4). Notably on DGM^4 and NewsCLIPpings, DREA improves by +2.9 and +4.3 points relative to DEFAME. This

Model	Instr-F1	mAP@50	mAP@75	Token-F1	Joint-F1
Qwen3-VL	55.40	45.20	30.10	40.50	35.20
GPT-4o	60.10	50.30	40.20	55.40	45.10
DEFAME	62.30	48.50	35.40	45.60	42.30
DREA (Ours)	**63.50**	**54.40**	**44.20**	**56.80**	**49.60**

Table 3: Fine-grained grounding on Falsdo (%).

Model	Falsdo	DGM[4]	NewsCLIPpings	FakeNewsNet
FKA-Owl	55.10	-	55.34	70.20
GPT-4o	67.34	60.50	73.10	82.50
DEFAME	70.63	72.30	74.20	84.10
Qwen3-VL	67.17	64.40	68.50	75.60
DREA (Ours)	**75.88**	**75.20**	**78.50**	**88.40**

Table 4: Cross-benchmark veracity classification (Acc, %).

Model	T-F/L	T-S/T	I-Temp	I-Obj
Qwen3-VL	45.20	48.50	42.10	40.30
GPT-4o	58.40	60.20	55.60	52.80
DEFAME	55.60	52.30	50.40	48.90
DREA (Ours)	**65.80**	**68.40**	**62.50**	**60.10**

Table 5: Accuracy (%) of manipulation type prediction on Falsdo. **T-F/L**: Text-Fact/Logic; **T-S/T**: Text-Style/Tone; **I-Temp**: Image-Temporal; **I-Obj**: Image-Object.

suggests that reliability-aware evidence auditing improves robustness against open-world evidence noise, validating the system's generalization capability beyond the specific distribution of FALSDO.

5.3 Fine-Grained Grounding

We evaluate the model's ability to output verifiable evidence chains. We report box localization quality using mAP@50 and mAP@75 following common object detection evaluation protocols [Yu *et al.*, 2016]. We report Joint-F1, which counts a prediction as correct *only if* both the verdict and the evidence (box/span) are accurate. Results in Table 3 show DREA's Joint-F1 reaches 49.60, a **+14.4** leap over the Base model and a **+4.5** lead over GPT-4o. This significant gap quantifies the "hallucination vs. grounding" difference: while monolithic models may guess the right label, DREA provides auditable proof, aligning with the benchmark's goal of trustworthy verification.

5.4 Manipulation Type Prediction

We evaluate the distinguishability of specific tampering patterns like Temporal Displacement, Semantic Conflict. DREA leads significantly in all categories (Table 5), especially in "Text Fact/Logic" +20.6 and "Image Temporal Displacement" +20.4. This confirms that decoupling evidence channels allows for sharper diagnosis of complex causal and temporal contradictions compared to entangled monolithic representations. In other words, DREA is explicitly "checking the story," not just "reading the pixels."

Variant	Acc	AUC	Gain
Qwen3-VL (Base)	67.17	67.18	-
Single-Channel Verifiers:			
Narrative Fact Verifier Only	69.85	69.10	+2.68
Cross-Modal Verifier Only	65.60	64.80	-1.57
Scene Logic Verifier Only	62.15	61.20	-5.02
System Component Ablation:			
w/o Decoupled Auditing (Monolithic EAM)	73.42	72.81	+6.25
w/o Evidence Acquisition (Reasoning Only)	70.42	69.48	+3.25
Full Reference System (DREA)	**75.88**	**75.18**	**+8.71**

Table 6: Ablation of DREA components on Falsdo (%). Gain is computed w.r.t. the Base backbone.

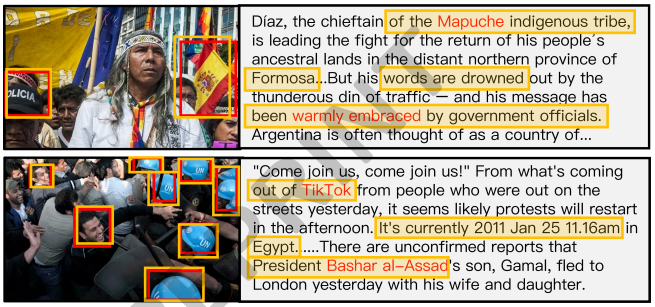


Figure 7: Visualization of detection and grounding results. Ground truth annotations are in red, and prediction results are in orange.

5.5 Ablation Study: Validating Decoupling

Table 6 shows complementarity rather than a single driver. Evidence acquisition determines *what* the system can reason over: removing it causes the larger regression (70.42 vs. 73.42 for monolithic auditing). Decoupling determines *how* conflicting evidence is used: the full system still gains 2.46 points over monolithic EAM and has a much smaller artifact-control drop. The weak Scene Logic Verifier confirms that FALSDO cannot be solved by visual artifacts alone.

6 Conclusion

This work presents FALSDO, an artifact-controlled diagnostic benchmark for grounded multimodal misinformation verification, and DREA, a reference evidence-auditing baseline for instantiating its protocols. Experiments show that artifact equalization exposes shortcut reliance, while reliability-aware auditing improves robustness and grounding. Future directions include broader generator sensitivity analysis and lower-latency deployment.

Ethical Statement

FALSDO is intended for research on misinformation detection and grounding. We do not release any automated tool for generating misinformation at scale, and we recommend downstream users follow institutional review and dataset governance practices. We further discourage any use that facilitates deception, manipulation, or targeted influence operations.

References

[Ali Adeeb and Mirhoseini, 2023] Rana Ali Adeeb and Mahdi Mirhoseini. The impact of affect on the perception of fake news on social media: a systematic review. *Social Sciences*, 12(12):674, 2023.

[Arjovsky *et al.*, 2020] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[Augenstein *et al.*, 2019] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, 2019.

[Boididou *et al.*, 2018] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86, 2018.

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

[Bui *et al.*, 2024] Hien Thu Bui, Viachaslau Filimonau, and Hakan Sezerel. Ai-thenticity: Exploring the effect of perceived authenticity of ai-generated visual content on tourist patronage intentions. *Journal of Destination Marketing & Management*, 34:100956, 2024.

[Chen *et al.*, 2022] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2897–2905, New York, NY, USA, 2022. Association for Computing Machinery.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[Jin *et al.*, 2017] 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*, pages 795–816, 2017.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017.

[Lakara *et al.*, 2025] Kumud Lakara, Georgia Channing, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder de Witt. MAD-sherlock: Multi-agent debate for visual misinformation detection. In *ICML 2025 Workshop on Collaborative and Federated Agentic Workflows*, 2025.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[Li *et al.*, 2025] Guoyi Li, Die Hu, Xiaomeng Fu, Qirui Tang, Yulei Wu, Xiaodan Zhang, and Honglei Lyu. Entity graph alignment and visual reasoning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2486–2495, 2025.

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

[Mao *et al.*, 2016] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Mirsky and Lee, 2021] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), January 2021.

[Nakamura *et al.*, 2020] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille,

France, May 2020. European Language Resources Association.

[Newman and Schwarz, 2024] Eryn J Newman and Norbert Schwarz. Misinformed by images: How images influence perceptions of truth and what can be done about it. *Current Opinion in Psychology*, 56:101778, 2024.

[Papadopoulos *et al.*, 2024] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.

[Pulm *et al.*, 2024] Claudine Pulm, Anne Gast, and Jan Rummel. A picture corrects a thousand words–the effect of photos on veracity feedback. *Consciousness and Cognition*, 125:103758, 2024.

[Qian *et al.*, 2021] Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(3), July 2021.

[Radford *et al.*, 2021] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[Sagawa *et al.*, 2020] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.

[Schlichtkrull *et al.*, 2023] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167, 2023.

[Shao *et al.*, 2023] 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*, pages 6904–6913, 2023.

[Shopnil *et al.*, 2025] Mir Nafis Sharear Shopnil, Sharad Duwal, Abhishek Tyagi, and Adiba Mahbub Proma. Mirage: Agentic framework for multimodal misinformation detection with web-grounded reasoning. *arXiv preprint arXiv:2510.17590*, 2025.

[Shu *et al.*, 2020] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

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

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification, 2018.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[Wang *et al.*, 2018] 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 & data mining*, pages 849–857, 2018.

[Wang, 2017] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection, 2017.

[Wilson *et al.*, 2023] Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12):230964, 2023.

[Wu *et al.*, 2021] 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*, pages 2560–2569, 2021.

[Wu *et al.*, 2025] Yin Wu, Zhengxuan Zhang, Fuling Wang, Yuyu Luo, Hui Xiong, and Nan Tang. Exclaim: An explainable cross-modal agentic system for misinformation detection with hierarchical retrieval, 2025.

[Yang *et al.*, 2025] Shuo Yang, Zijian Yu, Zhenzhe Ying, Yuqin Dai, Guoqing Wang, Jun Lan, Jinfeng Xu, Jinze Li, and Edith C. H. Ngai. Rama: Retrieval-augmented multi-agent framework for misinformation detection in multi-modal fact-checking, 2025.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 69–85, Cham, 2016. Springer International Publishing.

[Yu *et al.*, 2024] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

[Yue *et al.*, 2024] Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. Evidence-driven retrieval augmented response generation for online misinformation, 2024.