RETRACE: REASONING THROUGH VERIFICATION STRATEGIES AND EMOTION-GUIDED EVIDENCE REPAIR FOR MULTIMODAL FACT-CHECKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective multimodal fact-checking requires more than retrieving topically relevant information: a system must determine which investigative route is most likely to yield decisive evidence and recognize when the retrieved evidence leaves important aspects of a claim unresolved. We present RETRACE (REasoning through verification sTRAtegies and emotion-guided evidenCE repair), a dual-stage adaptive framework for evidence-grounded multimodal fact-checking. RETRACE first generates multiple claim-specific verification strategies representing alternative paths to decisive evidence and ranks them through pairwise preference modeling. The selected strategy is then translated into prioritized, executable forensic tasks and carried out with multimodal tools to construct an initial evidence state. After observing the initial evidence, it diagnoses unresolved evidential gaps by jointly assessing factual evidence coverage with textual, visual and cross-modal emotional signals to repair the evidence state. It initiates a targeted second investigation aimed at recovering the missing or insufficient evidence. An independent evidence-grounded verifier then determines the final claim label from the evidence state. Without task-specific fine-tuning, RETRACE achieves accuracies of 80.8% on AVeriTeC, 87.1% on MOCHEG, 86.7% on VERITE, and 84.2% on MMFakeBench, improving over the strongest reported comparison on each benchmark by 2.8, 25.9, 2.8, and 8.9 percentage points, respectively. An ablation study further demonstrates that verification strategy discovery and emotion-guided evidence repair provide complementary benefits to fact-checking.

1 INTRODUCTION

Online misinformation remains a persistent challenge for digital information ecosystems, while advances in generative models continue to diversify how misleading content is produced and disseminated. Recent studies highlight both the difficulty of characterizing misinformation harms and the value of professional fact-checking in correcting false beliefs (Budak et al., 2024; Martel & Rand, 2024). Increasingly, misinformation is *multimodal*: misleadingness may originate from textual distortion, manipulated visual content, or an incorrect relationship between otherwise legitimate text and images (Van Der Meer et al., 2025; Liu et al., 2025; Papadopoulos et al., 2024; Yao et al., 2023a). Fact-checking has moved toward *evidence-grounded verification*, with benchmarks such as AVeriTeC, MOCHEG, and AVerImaTeC requiring systems to retrieve external evidence rather than rely solely on parametric knowledge (Yao et al., 2023a; Schlichtkrull et al., 2024; Cao et al., 2026).

Recent fact-checking systems improve evidence acquisition by incorporating external context, multi-document evidence, multimodal consistency, and knowledge augmentation (Qi et al., 2024; Chen et al., 2024b; Ma et al., 2024; Yan et al., 2025; Hu et al., 2025), while DEFAME and subsequent evidence-grounded approaches demonstrate the effectiveness of dynamic multimodal retrieval and web evidence (Braun et al., 2025; Duwal et al., 2026). Yet retrieving evidence does not by itself determine *how a particular claim should be investigated* or *whether the acquired evidence is sufficient*.

Two limitations remain particularly important. **First**, existing systems reason primarily about verification operations rather than verification strategy. LLM-based fact-checkers and general-purpose agents can decompose claims, formulate queries, plan retrieval, select tools, synthesize evidence, and iteratively refine searches (Pan et al., 2023; Li et al., 2024; Zhao et al., 2024; Khaliq et al.,

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.

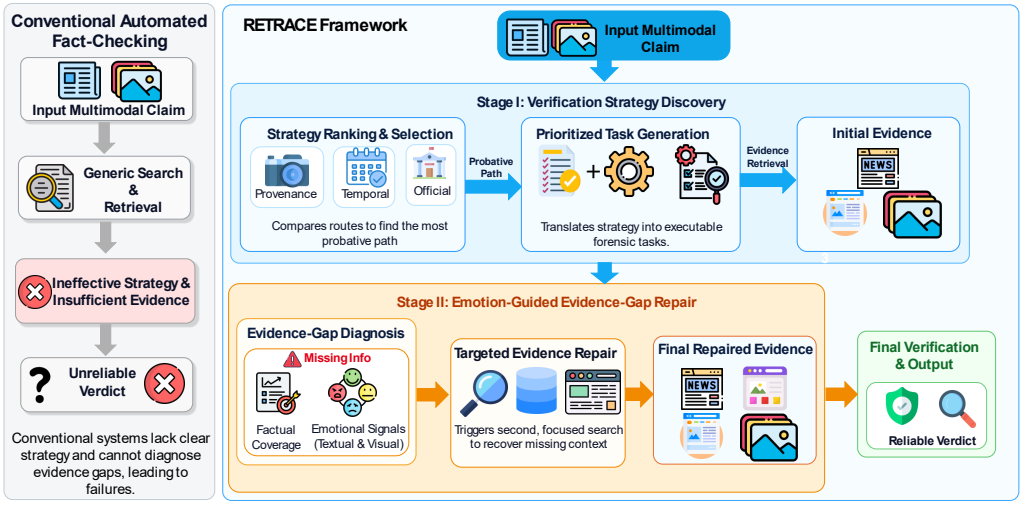


Figure 1: Overview of RETRACE versus conventional evidence-grounded verification. RETRACE adaptively selects verification strategies, identifies unresolved evidence gaps, performs targeted repair, and uses an independent verifier to produce the final verdict.

2024; Yue et al., 2024; Rothermel et al., 2024; Chowdhury et al., 2025; Tariq & Kementchedjhieva, 2026; Yao et al., 2023b). However, they do not explicitly compare alternative investigative strategies before retrieval. Consequently, they may select the next query or action without determining whether the claim is best verified through provenance tracing, temporal validation, source attribution, official-record corroboration, document identification, or manipulation analysis.

Second, even a well-planned investigation may retrieve evidence that is relevant but still insufficient. Topical relevance does not guarantee evidential utility, as retrieved evidence may be incomplete, redundant, weakly supportive, or affected by misleading contextual cues (Zheng et al., 2024; Wan et al., 2024; Li et al., 2025; Dang et al., 2025). Repeated retrieval can gather more information without identifying *what remains unresolved*. We therefore use multimodal emotional framing as a complementary diagnostic signal: emotion may differ across textual and visual modalities (Wu et al., 2025; Bagdon et al., 2025) and has previously been used for misinformation prediction (Ma et al., 2024). RETRACE uses emotion to prioritize salient claim components whose factual evidence remains insufficient, complementing evidence coverage in verification (Fatahi Bayat et al., 2025).

To address these limitations, we introduce **RETRACE** (**RE**asoning through verification s**TRA**tegies and emotion-guided eviden**CE** repair), a dual-stage adaptive framework for evidence-grounded multimodal fact-checking (Figure 1). Before retrieval, RETRACE performs *verification strategy discovery* by generating claim-specific investigative directions, ranking them through pairwise comparisons, and converting the selected strategy into prioritized forensic tasks. After initial evidence acquisition, it performs *emotion-guided evidence-gap repair*, jointly using factual claim–evidence coverage and emotional signals to identify unresolved evidence and guide focused follow-up retrieval. An evidence-grounded verifier then predicts the label from the repaired evidence state.

Our contributions are threefold. **First**, we introduce *verification strategy discovery*, which reasons over and competitively ranks alternative investigative directions before converting the selected strategy into executable forensic tasks. **Second**, we introduce *emotion-guided evidence-gap repair*, which combines factual evidence coverage with multimodal emotional framing to identify unresolved evidence and guide targeted follow-up retrieval, without using emotion as a direct veracity cue. **Third**, we systematically evaluate RETRACE across four textual and multimodal fact-checking benchmarks, with controlled ablations demonstrating the individual effects of its two stages.

2 Related Works

Automated Fact-Checking (AFC) spans claim analysis, evidence retrieval, verification, and explanation (Akhtar et al., 2023; Wang et al., 2024; Schlichtkrull et al., 2024; Zerong et al., 2025). RETRACE builds on recent evidence-grounded AFC but focuses on two complementary control

problems: selecting an effective verification strategy before retrieval and diagnosing what evidence remains unresolved after retrieval. Table 1 summarizes the closest systems.

Evidence-grounded fact-checking. Recent LLM-based AFC systems move beyond direct prediction from parametric knowledge by integrating explicit evidence acquisition with structured reasoning. PROGRAMFC generates executable reasoning programs (Pan et al., 2023), SELF-CHECKER modularizes verification (Li et al., 2024), PACAR combines planning with customized actions (Zhao et al., 2024), and open-web systems decompose complex claims before retrieving temporally appropriate evidence (Chen et al., 2024a). InFact structures verification as a multi-stage retrieval pipeline (Rothermel et al., 2024), while RAV and related reranking methods emphasize the importance of evidence selection and ranking (Zheng et al., 2024; Malviya & Katsigiannis, 2024). RAFTS synthesizes contrastive arguments from retrieved evidence (Yue et al., 2024), and complementary work studies evidence-conditioned LLM verification, efficient grounding, fine-grained decomposition, question-driven verification, and long-form factuality assessment (Guan et al., 2024; Tang et al., 2024b; Mitra et al., 2025; Chowdhury et al., 2025; Wan et al., 2025). These methods improve evidence retrieval and reasoning, while RETRACE selects the best investigative strategy that guides retrieval.

Multimodal fact-checking. Multimodal fact-checking has evolved from detecting content manipulation and image–text inconsistency toward evidence-grounded verification with external context. MOCHEG introduced multimodal evidence retrieval and verification (Yao et al., 2023a), while HAMMER and causal-debiasing methods address fine-grained manipulation and modality-specific shortcuts (Shao et al., 2023; Chen et al., 2023). Subsequent work explores external-context retrieval, image meta-context recovery, multimodal evidence summarization, document inconsistency, synthetic-data transfer, event-level inconsistency, chart verification, and social-media reasoning (Qi et al., 2024; Tonglet et al., 2024; Chen et al., 2024b; Tang et al., 2024a; Zeng et al., 2024; Ma et al., 2024; Akhtar et al., 2024; Jin et al., 2024), with COVE emphasizing out-of-context image recovery (Tonglet et al., 2025). Recent MLLM and agentic systems further incorporate knowledge augmentation, multimodal tools, iterative retrieval, multi-agent collaboration, and modality-selective evidence acquisition (Liu et al., 2025; Yan et al., 2025; Hu et al., 2025; Khaliq et al., 2024; Braun et al., 2025; Cao et al., 2026; Jung et al., 2026b; Tariq & Kementchedjhieva, 2026; Tsoneva et al., 2026; Duwal et al., 2026). RETRACE differs by explicitly separating verification-strategy selection before investigation from evidence-gap repair after initial retrieval

Agentic planning and verification strategies. Deliberative language agents improve complex task solving by exploring alternative reasoning trajectories and adapting their actions through environmental feedback. TREE OF THOUGHTS searches over candidate reasoning paths (Yao et al., 2023b), while LATS integrates search, action, feedback, and reflection (Zhou et al., 2024). Tool-oriented agents extend these capabilities through hierarchical tool retrieval, tool-path planning, aligned invocation, and cooperative execution (Du et al., 2024; Wu et al., 2024; Chen et al., 2024c; Shen et al., 2024). These approaches primarily deliberate over individual reasoning steps, subproblems, or tool actions. RETRACE instead extends deliberation to a higher level of abstraction: claim-specific verification strategy selection. Before tool execution, it generates alternative investigative directions such as provenance tracing, temporal validation, source attribution, official-record corroboration, and manipulation analysis, and ranks them according to their expected evidential utility. The selected strategy is then instantiated as a set of prioritized, executable forensic tasks.

Emotion-aware multimodal reasoning. Emotion has previously been used as a feature for multimodal misinformation detection; for example, Event-Radar incorporates textual emotion alongside visual and cross-modal inconsistencies (Ma et al., 2024). More broadly, multimodal affective research examines emotion recognition, emotion-cause analysis, and modality-dependent emotional expression (Hu et al., 2024; Bagdon et al., 2025; Wu et al., 2025). However, emotional expression is not itself evidence of veracity, and MLLMs can exhibit modality-dependent biases when attributing emotion. RETRACE therefore uses emotion only as an auxiliary evidence-acquisition signal. Specifically, textual, visual, and cross-modal emotional signals reweight the priority of requirements already identified as insufficiently supported through factual coverage, thereby directing follow-up retrieval toward potentially salient evidence gaps. Emotion controls evidence acquisition but does not directly inform the final veracity prediction.

Table 1: **Comparison of representative automated fact-checking systems.** ✓ denotes an explicit capability, ○ a partial or implicit capability, and – an absent or not explicitly described capability.

Method	MM Claim	MM Evidence	Evid. Retrieval	Multi-step	Plan	Tools	Verify Strategy	Evidence Sufficiency	Targeted Repair	Emotion-guided	Explanation
ProgramFC (Pan et al., 2023)	–	–	✓	○	✓	✓	–	–	–	–	✓
Self-Checker (Li et al., 2024)	–	–	✓	–	○	–	–	–	–	–	○
PACAR (Zhao et al., 2024)	–	–	✓	✓	✓	✓	–	–	–	–	✓
RAGAR (Khaliq et al., 2024)	✓	–	✓	✓	○	–	–	–	○	–	✓
MOCHEG (Yao et al., 2023a)	–	✓	✓	–	–	–	–	–	–	–	✓
SNIFFER (Qi et al., 2024)	✓	–	✓	–	–	✓	–	–	–	–	✓
MetaSumPerceiver (Chen et al., 2024b)	–	✓	–	–	–	–	–	–	–	–	✓
Event-Radar (Ma et al., 2024)	✓	–	–	–	–	–	–	–	–	–	–
MMD-Agent (Liu et al., 2025)	✓	–	✓	–	✓	✓	–	–	–	–	✓
DEFAME (Braun et al., 2025)	✓	✓	✓	✓	✓	✓	–	✓	✓	–	✓
REVEAL (Tariq & Kementchedjhieva, 2026)	✓	–	✓	✓	○	–	–	○	○	–	✓
VILLAIN (Jung et al., 2026b)	✓	✓	✓	✓	○	–	–	–	–	–	✓
Selective-MM Retrieval (Tsoneva et al., 2026)	✓	–	✓	○	○	–	–	–	–	–	✓
Evidence-Grounded GNN (Duwal et al., 2026)	✓	–	✓	–	–	✓	–	–	–	–	–
RETRACE (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

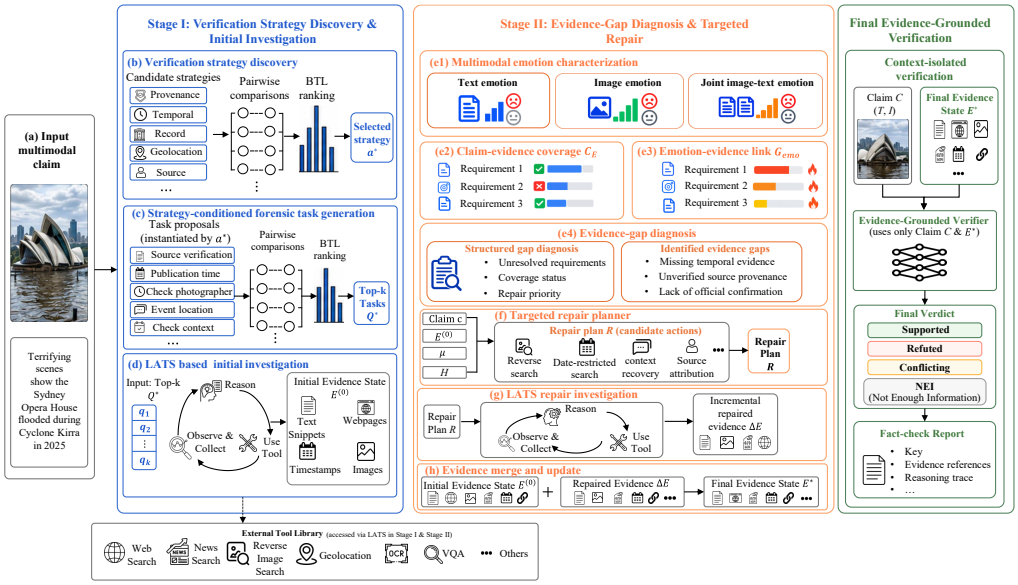


Figure 2: RETRACE architecture. Stage I determines **how to investigate** a claim by selecting a claim-specific verification strategy, prioritizing forensic tasks, and constructing the initial evidence state $E^{(0)}$. Stage II determines **what remains unresolved** by diagnosing evidence gaps from factual coverage and multimodal emotion signals and performing targeted repair to obtain E^*. A context-isolated verifier then predicts the final verdict using only the claim and the repaired evidence.

3 RETRACE

RETRACE is an evidence-grounded multimodal fact-checking framework that adapts its investigation at two distinct points. Given a textual or image–text claim $C = (T, I)$, RETRACE first determines *how the claim should be investigated* by discovering and ranking alternative verification strategies, translating the selected strategy into executable tasks, and collecting an initial evidence state. It then determines *what the acquired evidence still fails to resolve* through emotion-guided evidence-gap diagnosis and targeted evidence repair. An independent verifier predicts the label using the final evidence. Figure 2 provides an overview of the RETRACE Architecture.

The overall evidence flow can be summarized as

$$C \to a^* \to \mathcal{Q}^* \to E^{(0)} \to \mathcal{G} \to E^* \to \hat{y}, \tag{1}$$

where a^* is the selected verification strategy, $\mathcal{Q}^*$ the prioritized task set, $E^{(0)}$ the initial evidence state, $\mathcal{G}$ the structured evidence gap diagnosis, E^* the repaired evidence state, and $\hat{y}$ the prediction.

Adaptive Stage I: Verification Strategy Discovery and Initial Investigation. Stage I performs initial adaptation by determining how the claim should be investigated before any tool execution.

Given claim C, RETRACE generates a set of candidate verification strategies,

$$\mathcal{A}(C) = a_1, \ldots, a_N, \tag{2}$$

where each a_i represents a high-level investigative route, such as provenance tracing, temporal verification, source attribution, official-record checking, or manipulation analysis. The candidates are compared pairwise according to their LLM-judged predicted evidential utility and aggregated using a Bradley–Terry–Luce (BTL) model,

$$P(a_i \succ a_j) = \frac{\exp(\sigma_i)}{\exp(\sigma_i) + \exp(\sigma_j)}, \tag{3}$$

where σ_i denotes the latent preference score of strategy a_i estimated from the pairwise judgments. The strategy with the highest score is selected,

$$a^* = \arg \max_{a_i \in \mathcal{A}(C)} \sigma_i. \tag{4}$$

Generate and Rank Forensic Tasks. The selected strategy is transformed into forensic tasks,

$$\mathcal{Q}(C, a^*) = q_1, \ldots, q_M. \tag{5}$$

As executing the complete task pool would be redundant and costly, RETRACE conducts a second pairwise tournament in which tasks are compared according to reliability, feasibility, probative value, and redundancy and only the highest-ranked tasks are retained:

$$\mathcal{Q}^* = \mathrm{TopK}_{q \in \mathcal{Q}(C, a^*)} \, \sigma(q), \tag{6}$$

where $\sigma(q)$ denotes the latent preference score assigned to forensic task q.

Collect Initial Evidence. The prioritized tasks are executed by an autonomous evidence-acquisition module based on a LATS-style search process, allowing actions to adapt to intermediate observations rather than following a fixed sequence. The initial evidence state is

$$E^{(0)} = \bigcup_{q \in \mathcal{Q}^*} \mathrm{Execute}(q). \tag{7}$$

Adaptive Stage II: Emotion-Guided Evidence-Gap Repair. Stage II performs post-retrieval adaptation by determining *what the initial evidence still fails to establish*. Its inputs are the original claim $C = (T, I)$, the initial evidence state $E^{(0)}$, and the Stage-I investigation history $\mathcal{H}^{(0)}$. RETRACE identifies unresolved factual requirements, prioritizes them using factual coverage and multimodal emotional framing, and performs targeted evidence repair.

Factual requirement extraction. We first use a structured GPT-5.2 call to decompose the claim into a small set of independently verifiable factual requirements,

$$\mathcal{P}(C) = p_1, \ldots, p_R. \tag{8}$$

Each requirement is represented as

$$p_j = (\phi_j, \kappa_j, m_j), \tag{9}$$

where ϕ_j is the factual proposition, κ_j denotes its verification type, and $m_j \in T, I, TI$ specifies whether the requirement primarily concerns the text, image, or their cross-modal relationship.

Multiview emotion characterization. We characterize emotional framing independently across three views. For text, the pretrained Emotion-English DistilRoBERTa model produces the distribution $\rho_T(e)$; for the image, EmotionCLIP-V2 produces a visual emotion distribution $\rho_I(e)$ that is deterministically mapped to the same emotion inventory; and for the joint image–text view, GPT-5.2 receives (T, I) and returns non-negative emotion scores that are normalized to obtain $\rho_{TI}(e)$.

$$\mathcal{E} = \mathrm{anger}, \mathrm{disgust}, \mathrm{fear}, \mathrm{joy}, \mathrm{sadness}, \mathrm{surprise}, \mathrm{neutral}. \tag{10}$$

From the normalized distributions, we compute the non-neutral intensity of each view,

$$\eta_v = 1 - \rho_v(\mathrm{neutral}), \qquad v \in T, I, TI, \tag{11}$$

and the normalized text–image emotional discrepancy,

$$D_{\mathrm{emo}} = \frac{\mathrm{JSD}(\rho_T \| \rho_I)}{\ln 2}, \qquad D_{\mathrm{emo}} \in [0, 1]. \tag{12}$$

The cross-modal emotional amplification is defined as

$$A_{\mathrm{emo}} = \max! \left(0, \eta_{TI} - \frac{\eta_T + \eta_I}{2}\right). \tag{13}$$

Requirement-level evidence coverage. An LLM-based GPT-5.2 evidence evaluator compares each requirement p_j against the complete initial evidence state $E^{(0)}$ and returns

$$(c_j, d_j, \mathcal{S}_j), \tag{14}$$

where c_j measures how completely the available evidence resolves the requirement, d_j denotes the evidential direction, and $\mathcal{S}_j$ contains the identifiers of evidence supporting the coverage assessment.

$$c_j \in 0, 0.25, 0.50, 0.75, 1.0, \tag{15}$$

corresponding to *unaddressed*, *weakly addressed*, *partially resolved*, *strong but incomplete*, and *decisively resolved*, respectively. Although these values encode ordered evidence-resolution categories, RETRACE treats the scale as interval-valued when computing the repair priorities below.

Emotion-conditioned evidence-gap calculation. Each requirement inherits an emotional-salience score from its modality tag:

$$\psi_j = \begin{cases} \eta_T, & m_j = \mathrm{T}, \\ \eta_I, & m_j = \mathrm{I}, \\ \dfrac{\eta_{TI} + D_{\mathrm{emo}} + A_{\mathrm{emo}}}{3}, & m_j = \mathrm{TI}. \end{cases} \tag{16}$$

Hence, $\psi_j \in [0, 1]$ quantifies only how strongly requirement p_j is associated with the emotional framing of the multimodal claim.

We combine emotional salience with factual insufficiency to obtain the requirement-level repair priority,

$$g_j = (1 - c_j)\left[(1 - \lambda) + \lambda\psi_j\right], \qquad 0 \le \lambda \le 1. \tag{17}$$

Here, λ is a fixed mixing coefficient controlling the influence of emotional salience on the prioritization of unresolved factual requirements. Because

$$(1 - \lambda) + \lambda\psi_j \in [1 - \lambda, 1], \tag{18}$$

emotion does not increase a requirement's repair priority beyond its raw factual insufficiency $(1 - c_j)$. Instead, emotionally neutral requirements can be down-weighted, whereas emotionally salient requirements retain a larger proportion of their factual-gap priority. Thus, factual insufficiency remains the primary signal and emotion acts only as a bounded prioritization factor.

The complete structured evidence-gap diagnosis is retained as

$$\mathcal{G} = \{(p_j, c_j, d_j, \psi_j, g_j)\}_{j=1}^{R}. \tag{19}$$

Because $p_j = (\phi_j, \kappa_j, m_j)$ already contains both the verification type and modality tag, these quantities are not repeated in $\mathcal{G}$.

Targeted repair planning and execution. Incomplete requirements are ranked according to g_j, and the highest-priority unresolved requirements form

$$\mathcal{U} = \mathrm{TopK}_{p_j : c_j < 1}\, g_j. \tag{20}$$

If all requirements satisfy $c_j = 1$, Stage II terminates without additional retrieval. Otherwise, a structured GPT-5.2 repair planner receives the original claim C, the initial evidence state $E^{(0)}$, the selected unresolved requirements $\mathcal{U}$, and the Stage-I investigation history $\mathcal{H}^{(0)}$:

$$\mathcal{X} = \pi_{\mathrm{repair}}\left(C, E^{(0)}, \mathcal{U}, \mathcal{H}^{(0)}\right). \tag{21}$$

Here, $\mathcal{X}$ denotes the resulting set of gap-specific repair actions. The resulting actions are executed through a LATS-style Reason–Act–Observe loop. Let ΔE denote the evidence recovered during this repair process:

$$\Delta E = \bigcup_{x \in \mathcal{X}} \mathrm{Execute}_{\mathrm{LATS}}(x). \tag{22}$$

The final evidence state is then

$$E^{*} = E^{(0)} \cup \Delta E. \tag{23}$$

Table 2: **Comparison of RETRACE with prior state-of-the-art methods and direct MLLM baselines across four fact-checking benchmarks.** Best results are shown in **bold** and second-best directly comparable results are underlined.

Method	AVeriTeC Acc.	MMFakeBench Acc.	MOCHEG Acc.	VERITE T/F Acc.
CFR (Sriram et al., 2024)	60.0	–	–	–
DeBERTa (Cao et al., 2023)	65.6	–	–	–
Papelo (Malon, 2024)	75.4	–	–	–
HerO (Yoon et al., 2024)	75.2	–	–	–
CAAFC (Eldifrawi et al., 2026)	78.0	–	–	–
LVLM4FV (Tahmasebi et al., 2024)	–	–	45.1	–
MetaSumPerceiver (Chen et al., 2024b)	–	–	48.6	–
AMuFC (Jung et al., 2026a)	–	–	61.2	–
CHASMA (Papadopoulos et al., 2024)	–	–	–	52.1
AITR (Papadopoulos et al., 2025)	–	–	–	58.0
DEFAME (Braun et al., 2025)	70.5	–	59.2	83.9
MMD-Agent (Liu et al., 2025)	–	62.1	–	–
T^2Agent (Cui et al., 2026)	–	75.3	–	–
GPT-4o	62.1	58.9	53.7	78.7
GPT-4o CoT	61.0	68.7	49.7	80.0
GPT-5.2	57.0	53.9	41.8	69.8
GPT-5.2 CoT	57.6	57.3	43.6	71.7
RETRACE (Ours)	**80.8**	**84.2**	**87.1**	**86.7**

Final Evidence-Grounded Verification. RETRACE transfers the final evidence state to a context-isolated verifier. For dataset D with label space $\mathcal{Y}_D$,

$$\hat{y} = J_D\left(C, E^*; \mathcal{Y}_D\right). \tag{24}$$

The verifier evaluates the original claim against the collected evidence, considers supporting and contradictory information to produce verdicts, and produces a concise evidence-backed explanation.

4 Experiments

Datasets. We evaluate RETRACE on four complementary fact-checking benchmarks. **AVeriTeC** is a real-world evidence-grounded fact-checking benchmark covering *Supported*, *Refuted*, *Not Enough Evidence (NEI)*, and *Conflicting Evidence/Cherrypicking (C/CP)* labels; we follow prior work and report four-class accuracy. **MOCHEG** evaluates end-to-end multimodal fact-checking using textual claims together with textual and visual evidence, and we report three-class accuracy over *Supported*, *Refuted*, and *NEI*. **VERITE** focuses on image–text out-of-context misinformation while reducing unimodal dataset bias; following the standard protocol, we report *True vs. OOC* (T/OOC), *True vs. Miscaptioned* (T/MC), and merged *True vs. False* (T/F) accuracy. **MMFakeBench** evaluates textual-veracity distortion (TVD), visual-veracity distortion (VVD), cross-modal consistency distortion (CCD), and real image–text pairs; following prior work, we evaluate on the official validation split and report macro-F1 and multiclass accuracy.

Model and Configuration. We use **GPT-5.2** as the fixed RETRACE backbone without task-specific fine-tuning. Verification strategies and forensic tasks are ranked via pairwise comparisons with BTL aggregation, then executed using multimodal retrieval and forensic tools. Stage II combines factual evidence coverage with multimodal emotion signals to trigger targeted repair of unresolved gaps. Strategy- and emotion-derived signals guide evidence acquisition only and are not passed to the final verifier; configurations are otherwise fixed across datasets except for label spaces and output formats.

Comparison to the State-of-the-Art In Table 2, we compare RETRACE with prior state-of-the-art methods and direct GPT-4o/GPT-5.2 baselines, with and without CoT. We report accuracy for AVeriTeC, MMFakeBench, MOCHEG, and T/F accuracy for VERITE. RETRACE achieves 80.8%, 84.2%, 87.1%, and 86.7% on AVeriTeC, MMFakeBench, MOCHEG, and VERITE, exceeding the

Table 3: Comparison of Stage-I investigation-planning mechanisms under the same backbone, tools, evidence-acquisition, and verification setting.

Planning Mechanism	**AVeriTeC** Acc.	**MOCHEG** Acc.	**VERITE** T/F Acc.	**MMFakeBench** Macro-F1
No explicit planning	68.6	75.2	75.0	73.5
Hierarchical Task Decomposition	72.4	79.6	78.8	78.0
Dynamic Tool-Action Planning	73.2	80.4	80.5	79.1
Strategic Query Formulation	74.2	81.3	80.9	79.7
Iterative Strategy Planning	<u>74.8</u>	<u>81.8</u>	<u>81.5</u>	<u>80.1</u>
VSD (Ours)	**80.8**	**87.1**	**86.7**	**85.3**

Table 4: Analysis of evidence-gap repair. All variants use the same Stage-I verification strategy discovery. Generic retrieval and all repair variants use the same additional retrieval budget.

Repair Strategy	**AVeriTeC** Acc.	**MOCHEG** Acc.	**VERITE** T/F Acc.	**MMFakeBench** Macro-F1
No second retrieval	76.8	82.6	83.9	80.7
Generic second retrieval	77.4	83.3	84.2	81.4
Coverage-guided repair	78.8	85.1	85.4	83.2
Emotion-guided repair	78.0	84.4	84.9	82.6
Full EGR	**80.8**	**87.1**	**86.7**	**85.3**

strongest directly comparable prior methods by 2.8, 8.9, 25.9, and 2.8 percentage points, respectively. Although direct GPT-5.2 does not consistently outperform GPT-4o, RETRACE substantially improves over direct GPT-5.2 with the same backbone; we therefore treat this gap as the difference between direct prediction and the full evidence-grounded system, while Stage-I and Stage-II ablations more directly isolate the proposed mechanisms.

Analysis of Verification Strategy Discovery We examine whether Stage I gains arise specifically from verification strategy discovery rather than planning in general. Table 3 compares VSD with progressively structured alternatives from no explicit planning to task decomposition, dynamic tool-action planning, strategic query formulation, and iterative strategy planning while keeping the backbone, tools, evidence acquisition, and final verifier fixed. VSD performs best across all four benchmarks, improving over iterative strategy planning from 74.8% to 80.8% on AVeriTeC, 81.8% to 87.1% on MOCHEG, 81.5% to 86.7% on VERITE, and 80.1% to 85.3% Macro-F1 on MMFakeBench, suggesting that explicitly comparing high-level verification strategies before committing to forensic tasks provides benefits beyond additional planning alone.

Evidence-gap repair analysis Table 4 evaluates whether Stage-II gains come from targeted gap diagnosis rather than simply adding another retrieval round. Generic second retrieval yields only modest improvements over stopping after Stage I, whereas coverage-guided and emotion-guided repair perform consistently better, showing that unresolved evidence is more effectively addressed when follow-up retrieval is explicitly prioritized. Coverage-guided repair remains stronger than emotion-guided repair alone, consistent with our design in which emotion acts only as a secondary search-control signal rather than factual evidence. Combining both signals in Full EGR gives the best result on all benchmarks such as 80.8% on AVeriTeC, 87.1% on MOCHEG, 86.7% T/F accuracy on VERITE, and 85.3% Macro-F1 on MMFakeBench.

Qualitative Analysis Figure 3 presents four representative examples illustrating RETRACE's verification process across different misinformation scenarios. For each case, we show the multimodal claim, the verification strategy selected in Stage I, the main evidence gap identified after the initial investigation, the decisive evidence recovered through targeted repair, and the final evidence-grounded verdict. The examples demonstrate how RETRACE adapts its investigation to different claims: Stage I determines the most informative verification route, while Stage II targets evidence that remains unresolved after the initial search. Importantly, the final verdict is supported by the retrieved evidence rather than by the intermediate strategy or emotion signals.



Figure 3: Representative RETRACE examples across datasets with predicted verdicts and evidence-grounded justifications.

Failure Analysis We manually analyze **350** incorrectly classified instances sampled across AVeriTeC, MOCHEG, VERITE, and MMFakeBench and identify five recurring failure modes. **(1) Label Ambiguity** arises when closely related benchmark categories are difficult to distinguish, such as *NEI* versus *Refuted* or *OOC* versus *Miscaptioned*. **(2) Evidence Access Failure** occurs when the appropriate investigation is selected but decisive evidence cannot be recovered because the original source is unavailable, poorly indexed, paywalled, archived incompletely, or contained only in video or other inaccessible media. **(3) Strategy–Task Mismatch** occurs when Stage I selects a plausible verification strategy but the resulting forensic tasks do not target the most probative evidence for the claim. **(4) Incomplete Gap Repair** captures cases where Stage II correctly identifies an unresolved requirement, such as provenance, date, location, or source identity, but the targeted follow-up investigation still fails to recover sufficiently decisive evidence. **(5) Verification Error** includes cases where adequate evidence is available but the final verifier misinterprets numerical, temporal, or cross-modal relations.

5 Conclusion

We presented RETRACE, a dual-stage adaptive framework for evidence-grounded multimodal fact-checking that explicitly controls evidence acquisition both before and after initial retrieval. RETRACE first discovers and ranks claim-specific verification strategies, translates the selected strategy into prioritized forensic tasks, and executes them through an adaptive tool-using investigation to construct an initial evidence state. It subsequently diagnoses unresolved factual requirements by combining evidence coverage with textual, visual, and cross-modal emotional signals, using emotion solely to prioritize evidence acquisition rather than as a veracity cue, and performs targeted evidence repair before final verification. Without task-specific fine-tuning, RETRACE achieves strong results across AVeriTeC, MOCHEG, VERITE, and MMFakeBench, demonstrating consistent effectiveness across textual and multimodal verification settings. Controlled analyses further show that verification-strategy discovery and evidence-gap repair provide complementary benefits beyond direct MLLM reasoning and non-adaptive retrieval. While challenges remain in open-web evidence quality, retrieval efficiency, and multimodal affect estimation, RETRACE provides a general framework for building more adaptive, evidence-grounded, and auditable multimodal fact-checking systems.

LLM USAGE STATEMENT

LLMs were used solely for minor grammatical editing of the manuscript. All methodological design, analysis, and writing were conducted under full human oversight and responsibility.

REFERENCES

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5430–5448, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.361.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. ChartCheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13921–13937, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.828.

Christopher Bagdon, Aidan Combs, Carina Silberer, and Roman Klinger. Donate or create? comparing data collection strategies for emotion-labeled multimodal social media posts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17307–17330, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.847.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. DEFAME: Dynamic evidence-based FAct-checking with multimodal experts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5383–5417. PMLR, 2025.

Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. Misunderstanding the harms of online misinformation. *Nature*, 630(8015):45–53, 2024. doi: 10.1038/s41586-024-07417-w.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. Are large language models good fact checkers: A preliminary study. *arXiv preprint arXiv:2311.17355*, 2023.

Rui Cao, Yulong Chen, Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. The automatic verification of image-text claims (AVerImaTeC) shared task. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pp. 74–90, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.fever-1.6.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3569–3587, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.196.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. MetaSumPerceiver: Multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8742–8757, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.474.

Zhi-Yuan Chen, Shiqi Shen, Guangyao Shen, Gong Zhi, Xu Chen, and Yankai Lin. Towards tool use alignment of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1382–1400, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.82.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 627–638, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.37.

Shayan Chowdhury, Sunny Fang, and Smaranda Muresan. FACT5: A novel benchmark and pipeline for nuanced fact-checking of complex statements. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pp. 101–117, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.fever-1.8.

Xing Cui, Yueying Zou, Zekun Li, Peipei Li, Xinyuan Xu, Xuannan Liu, and Huaibo Huang. T2agent: A tool-augmented multimodal misinformation detection agent with monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 175–183, 2026.

Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Jungang Li, Jingyu Wang, Peijie Jiang, Aiwei Liu, Jia Liu, and Xuming Hu. Exploring response uncertainty in MLLMs: An empirical evaluation under misleading scenarios. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18132–18173, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.916.

Yu Du, Fangyun Wei, and Hongyang Zhang. AnyTool: Self-reflective, hierarchical agents for large-scale API calls. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11812–11829. PMLR, 2024.

Sharad Duwal, Mir Nafis Sharear Shopnil, Abhishek Tyagi, and Adiba Mahbub Proma. Evidence-grounded multimodal misinformation detection with attention-based gnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10994–11003, June 2026.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. Caafc: Chronological actionable automated fact-checker for misinformation/non-factual hallucination detection and correction. *arXiv preprint arXiv:2605.12436*, 2026.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. FactBench: A dynamic benchmark for in-the-wild language model factuality evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33090–33110, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1587.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1090–1111, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.62.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. UniMEEC: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5248–5261, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.302.

Shuguo Hu, Jun Hu, and Huaiwen Zhang. Synergizing LLMs with global label propagation for multimodal fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1440, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.72.

Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. MM-SOC: Benchmarking multimodal large language models in social media platforms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6192–6210, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.370.

Jaeyoon Jung, Yejun Yoon, and Kunwoo Park. Is a picture worth a thousand words? adaptive multimodal fact-checking with visual evidence necessity. *arXiv preprint arXiv:2604.04692*, 2026a.

Jaeyoon Jung, Yejun Yoon, Seunghyun Yoon, and Kunwoo Park. VILLAIN at AVerImaTeC: Verifying image–text claims via multi-agent collaboration. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pp. 114–126, Rabat, Morocco, March 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.fever-1.9.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 280–296, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.29.

Dayang Li, Fanxiao Li, Bingbing Song, Li Tang, and Wei Zhou. IMRRF: Integrating multi-source retrieval and redundancy filtering for LLM-based fake news detection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9127–9142, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.461.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 163–181, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.12.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *International Conference on Learning Representations*, 2025.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.316.

Christopher Malon. Multi-hop evidence pursuit meets the web: Team papelo at fever 2024. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 27–36, 2024.

Shrikant Malviya and Stamos Katsigiannis. Evidence retrieval for fact verification using multi-stage reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7295–7308, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.428.

Cameron Martel and David G. Rand. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8:1957–1967, 2024. doi: 10.1038/s41562-024-01973-x.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. FactLens: Benchmarking fine-grained fact verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18085–18096, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.929.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6981–7004, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.386.

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. doi: 10.1007/s13735-023-00312-6.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5041–5050. IEEE, 2025.

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 (CVPR)*, pp. 13052–13062, June 2024. doi: 10.1109/CVPR52733.2024.01240.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. InFact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 108–112, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.12.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 1–26, Miami, Florida, USA, nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.fever-1.1.

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 (CVPR)*, pp. 6904–6913, June 2023.

Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang, and Fei Huang. Small LLMs are weak tool learners: A multi-LLM agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16658–16680, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.929.

Aniruddh Sriram, Fangyuan Xu, Eunsol Choi, and Greg Durrett. Contrastive learning to improve retrieval for real-world fact checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 264–279, 2024.

Sahar Tahmasebi, Eric Müller-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.

Chia-Wei Tang, Ting-Chih Chen, Kiet A. Nguyen, Kazi Sajeed Mehrab, Alvi Md Ishmam, and Chris Thomas. M3D: MultiModal MultiDocument fine-grained inconsistency detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22270–22293, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1243.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499.

Amina Tariq and Yova Kementchedjhieva. REVEAL: Retrieval-enhanced verification for multimodal fact-checking. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pp. 108–113, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.fever-1.8.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. "image, tell me your story!" predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7845–7864, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.448.

Jonathan Tonglet, Gabriel Thiem, and Iryna Gurevych. COVE: COntext and VEracity prediction for out-of-context images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2029–2049, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.102.

Yoana Tsoneva, Paul-Conrad Feig, Jiaao Li, Veronika Solopova, Neda Foroutan, Arthur Hilbert, and Vera Schmitt. Selective multimodal retrieval for automated verification of image–text claims. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pp. 127–135, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.fever-1.10.

Michiel Van Der Meer, Pavel Korshunov, Sébastien Marcel, and Lonneke van der Plas. HintsOfTruth: A multimodal checkworthiness detection dataset with real and synthetic claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31274–31291, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1510.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7468–7484, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.403.

Yingjia Wan, Haochen Tan, Xiao Zhu, Xinyu Zhou, Zhiwei Li, Qingsong Lv, Changxuan Sun, Jiaqi Zeng, Yi Xu, Jianqiao Lu, Yinhong Liu, and Zhijiang Guo. FaStFact: Faster, stronger long-form factuality evaluations in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23814–23854, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1295.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14199–14230, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.830.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6257–6274, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.332.

Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. ToolPlanner: A tool augmented LLM for multi granularity instructions with path planning and feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18315–18339, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1018.

Zehong Yan, Peng Qi, Wynne Hsu, and Mong-Li Lee. TRUST-VL: An explainable news assistant for general multimodal misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5588–5604, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.284.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743. Association for Computing Machinery, 2023a. doi: 10.1145/3539618.3591879.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023b.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. Hero at averitec: The herd of open large language models for verifying real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 130–136, 2024.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10331–10343, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.556.

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, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.613.

Zhaxi Zerong, Chenxi Li, Xinyi Liu, Ju-hui Chen, and Fei Xia. A systematic survey of claim verification: Corpora, systems, and case studies. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21452–21474, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1170.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12564–12573, Torino, Italia, May 2024. ELRA and ICCL.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9274–9281, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.551.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62138–62160. PMLR, 2024.

A Usage and Computational Cost

As GPT-5.2 and GPT-4o are accessed through OpenAI's API, the majority of MLLM inference is performed externally. On our side, the pretrained text- and image-emotion models used in Stage II (Emotion-English DistilRoBERTa and EmotionCLIP-V2) are executed locally on RTX 3090, while evidence processing, tool orchestration, BTL aggregation, and retrieval management are performed on Ubuntu. Since VERITE contains image–text claims throughout and therefore exercises both the multimodal investigation and emotion-guided repair components, we use VERITE as a representative multimodal setting and report average per-claim usage statistics in Table 5.

As expected, the direct GPT-4o and GPT-5.2 baselines require substantially fewer resources because they perform neither external evidence acquisition nor adaptive investigation. RETRACE requires 4:31 minutes per claim on average, with 18.6 MLLM calls, 6.8 search queries, approximately 63K input tokens, and 4.1K generated tokens. The corresponding average LLM/API cost is $0.18 per claim, while search API costs remain below $0.01.

The ablations further reveal where the additional computation is spent. Removing verification strategy discovery reduces the average runtime from 4:31 to 3:42 and lowers the number of MLLM calls from 18.6 to 13.2, reflecting the cost of strategy generation and pairwise preference aggregation. Removing evidence-gap repair produces a larger reduction, to 3:06 minutes and 4.2 searches per claim, because the second targeted investigation is entirely omitted. In contrast, generic second retrieval retains approximately the same search budget as RETRACE (6.8 searches per claim) and therefore has a similar runtime and API cost. This comparison is particularly useful for separating the computational effect of performing additional retrieval from the effect of RETRACE's gap-conditioned repair mechanism.

Coverage-only and emotion-only repair are moderately cheaper than the complete Stage-II mechanism because each omits one component of the gap-diagnosis process. Similarly, removing BTL-based task ranking reduces both MLLM calls and token processing, although it retains most of the

external evidence-acquisition cost. These efficiency results should therefore be interpreted together with the corresponding accuracy ablations: cheaper variants reduce investigation depth or decision control, whereas the complete RETRACE configuration incurs additional computation in exchange for stronger evidence-grounded verification.

Overall, RETRACE is substantially more expensive than direct MLLM prompting, but its additional cost arises from explicit evidence acquisition, strategy comparison, and targeted evidence repair rather than repeated unconstrained generation. At an average LLM/API cost of $0.18 per claim and search cost below $0.01, RETRACE remains practical for moderate-scale automated verification and can serve as an evidence-acquisition assistant for human fact-checkers. Scaling to high-volume social-media streams would nevertheless require upstream check-worthiness filtering or other claim-prioritization mechanisms.

Table 5: Usage and computational cost statistics of RETRACE, its ablation variants, and direct MLLM baselines. Values are reported per claim for Verite Dataset.

Group	Method	Time / Claim	LLM Calls	LLM/API Cost ($)	Search Cost ($)	# Search Queries	Input Tokens	Output Tokens
RETRACE	GPT-5.2	4:31	18.6	0.18	< 0.01	6.8	63K	4.1K
Ablations	w/o Strategy Discovery	3:42	13.2	0.14	< 0.01	6.1	49K	3.4K
	w/o Evidence-Gap Repair	3:06	12.4	0.13	< 0.01	4.2	46K	3.0K
	Generic Second Retrieval	4:19	15.1	0.16	< 0.01	6.8	56K	3.6K
	Coverage-Only Repair	4:08	15.7	0.16	< 0.01	6.4	55K	3.6K
	Emotion-Only Repair	3:57	14.9	0.15	< 0.01	6.0	52K	3.5K
	w/o BTL Task Ranking	3:51	14.0	0.15	< 0.01	6.5	51K	3.4K
Baselines	GPT-4o	0:08	1	0.01	0	0	2.4K	0.18K
	GPT-4o CoT	0:14	1	0.02	0	0	3.1K	0.62K
	GPT-5.2	0:10	1	0.01	0	0	2.5K	0.21K
	GPT-5.2 CoT	0:17	1	0.02	0	0	3.3K	0.71K

B Robustness to Fact-Checking Domain Exclusion

To assess whether RETRACE benefits from directly retrieving existing professional fact-checking articles, we repeat the evaluation after excluding major fact-checking domains from all web-search results. Any retrieved URL matching an excluded domain is removed before it can be accessed by the investigation agent. The remainder of the pipeline, including the GPT-5.2 backbone, verification-strategy discovery, retrieval budget, evidence-gap repair, and final verifier, is kept unchanged.

Table 6: Robustness of RETRACE when professional fact-checking domains are excluded from web retrieval.

Retrieval Setting	**AVeriTeC** Acc.	**MMFakeBench** Acc.	**MOCHEG** Acc.	**VERITE** T/F Acc.
Full Web	80.8	84.2	87.1	86.7
w/o Fact-Checking Domains	78.8	82.9	75.4	85.4
Δ	-2.0	-1.3	-11.7	-1.3

Table 6 evaluates RETRACE under a stricter retrieval setting in which professional fact-checking websites are unavailable. Performance remains comparatively stable on AVeriTeC, MMFakeBench, and VERITE, indicating that the framework can recover useful evidence from independent web sources rather than relying on existing fact-check verdicts. The larger degradation on MOCHEG suggests a stronger dependence of its evidence environment on professionally fact-checked claim sources; nevertheless, RETRACE retains substantial performance after those domains are removed.

Table 7: Fact-checking domains excluded from retrieval in the leakage-controlled evaluation.

Excluded Fact-Checking Domains / URLs	
snopes.com	politifact.com
factcheck.org	truthorfiction.com
fullfact.org	leadstories.com
hoax-slayer.net	checkyourfact.com
reuters.com/fact-check	reuters.com/article/fact-check
apnews.com/APFactCheck	factcheck.afp.com
poynter.org	factcheck.ge
vishvasnews.com	boomlive.in
altnews.in	thequint.com/news/webqoof
factcheck.kz	–

This experiment provides an explicit safeguard against label leakage through directly retrieved fact-check pages.

C RETRACE CONFUSIONS

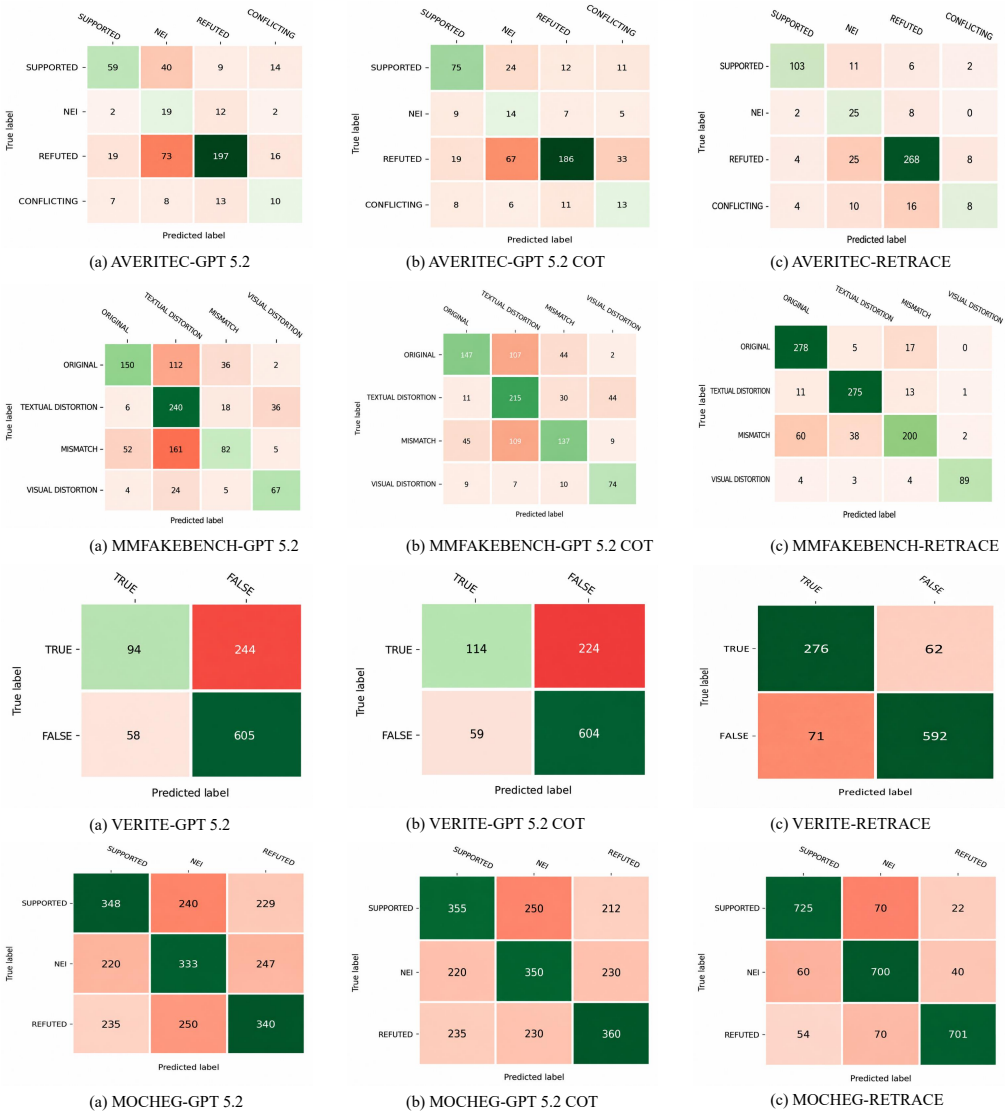
Figure 4 provides a class-level comparison of direct GPT-5.2, GPT-5.2 with chain-of-thought prompting, and RETRACE across the four benchmarks. Across all four benchmarks, RETRACE produces a much stronger diagonal concentration than direct GPT-5.2 and GPT-5.2 CoT, indicating more consistent class-wise discrimination. On AVeriTeC, RETRACE substantially improves Supported, NEI, and Refuted recognition, although Conflicting Evidence remains difficult and is often confused with Refuted.

On MMFakeBench, the largest gain appears in the Mismatch category, where direct GPT-5.2 frequently predicts Textual Distortion instead. RETRACE markedly reduces this confusion while also improving Original, Textual Distortion, and Visual Distortion recognition, suggesting stronger cross-modal verification.

On VERITE and MOCHEG, RETRACE also reduces major baseline biases. It substantially decreases True-to-False errors on VERITE and achieves balanced recognition of Supported, NEI, and Refuted classes on MOCHEG, showing that the overall gains are distributed across classes rather than driven by a single dominant label.

D DISCUSSION

Although RETRACE is designed to operate with different MLLM backbones, the quality of strategy generation, evidence assessment, repair planning, and final verification remains dependent on the capabilities of the underlying model. Its modular design nevertheless allows the reasoning backbone, emotion encoders, retrieval tools, and forensic modules to be replaced independently as stronger components become available. Despite the consistent improvements observed across our benchmarks, RETRACE has several limitations. **(i) Credibility of External Evidence**: RETRACE relies on open-web and multimodal retrieval, where retrieved sources may be incomplete, outdated, biased, or mutually inconsistent. This is particularly important for fact-checking because semantically relevant evidence is not necessarily trustworthy or decisive. We mitigate this risk by preserving source and provenance information, favoring original or authoritative sources during investigation, and seeking corroborating evidence when possible. Explicit source-quality calibration and credibility-aware retrieval remain promising directions for improving verification reliability. **(ii)System Stability and Sensitivity:** The investigation process depends on several stochastic and external components, including MLLM-based pairwise comparisons, structured evidence assessment, search-engine responses, reverse-image retrieval, and webpage accessibility. Consequently, small variations in model outputs or tool availability can alter the selected strategy or evidence trajectory. BTL aggregation and structured intermediate representations reduce sensitivity to individual judgments, but do not eliminate it. Improving ranking stability, retrieval reproducibility, and robustness to unavailable external tools is therefore an important direction for future work. **(iii)Emotion Guidance and Model Reliability:** RETRACE uses emotion only as a secondary sig-



(a) AVERITEC-GPT 5.2 (b) AVERITEC-GPT 5.2 COT (c) AVERITEC-RETRACE

(a) MMFAKEBENCH-GPT 5.2 (b) MMFAKEBENCH-GPT 5.2 COT (c) MMFAKEBENCH-RETRACE

(a) VERITE-GPT 5.2 (b) VERITE-GPT 5.2 COT (c) VERITE-RETRACE

(a) MOCHEG-GPT 5.2 (b) MOCHEG-GPT 5.2 COT (c) MOCHEG-RETRACE

Figure 4: Confusion matrices of Direct GPT-5.2, GPT-5.2 CoT, and RETRACE across AVeriTeC, MMFakeBench, VERITE, and MOCHEG. RETRACE shows stronger diagonal concentration and reduced class confusion across datasets.

nal for prioritizing unresolved evidence and explicitly excludes emotion variables from the final verifier. Nevertheless, the pretrained textual and visual emotion models, as well as GPT-5.2 used for joint multimodal emotion characterization, may exhibit contextual or demographic biases that can affect which evidence receives additional scrutiny. In addition, hallucination remains possible in requirement decomposition, repair planning, and evidence interpretation because these components rely on MLLM reasoning. Future work should investigate emotion calibration, bias-aware repair policies, and stronger verification of model-generated intermediate decisions.

E CASE STUDY

Figures 5–12 present two representative RETRACE fact-checking case studies with correct final predictions. Figures 5–8 illustrate the first case, while Figures 9–12 show a second example. In each case, Stage I first generates and ranks claim-specific verification strategies, selects the most suitable investigative direction, prioritizes the corresponding forensic tasks, and executes evidence-seeking actions to construct the initial evidence state $E^{(0)}$. Stage II then examines the acquired evidence at the requirement level, identifies insufficiently resolved factual components using evidence coverage together with multimodal emotion-guided prioritization, and performs targeted evidence repair to recover the missing or weakly supported information. The resulting repaired evidence state E^{*} is finally passed, together with the original claim, to the context-isolated verifier, which produces the final evidence-grounded verdict and supporting explanation. These examples demonstrate how RETRACE adapts both the initial investigation strategy and the subsequent repair process according to the specific evidential needs of each claim.



News Text: A shark was photographed swimming on a flooded highway in Houston, Texas, during Hurricane Harvey in August 2017.
GT: REFUTED
PRED: REFUTED

Claim Report

The claim associates an image of a shark swimming beside a vehicle on a flooded road with Hurricane Harvey in Houston. Hurricane Harvey genuinely caused catastrophic flooding in southeastern Texas, including the Houston metropolitan area. However, provenance investigation shows that the shark in the viral image originates from an earlier photograph by Thomas P. Peschak of a great white shark following a kayak off South Africa. Peschak's photograph was repeatedly reused in digitally manipulated disaster images, and National Geographic specifically documents its reappearance in a fake Houston flood image during Hurricane Harvey. Thus, the underlying flood event is authentic, but the image presented as evidence of a shark swimming on a Houston highway is manipulated .

References:
[1] Thomas P. Peschak, White Shark Kayak / photograph provenance.
[2] National Geographic, How the ultimate shark photo went viral.
[3] Thomas P. Peschak, National Geographic, My shark photo took over the internet.

========== RETRACE DECISION TRACE ==========
========== STAGE I: VERIFICATION STRATEGY DISCOVERY ==========

[ASSAY PIPELINE] Starting investigation.

[RETRACE] Generating candidate verification assays.

[RETRACE] Ranking assays with pairwise comparison + BTL aggregation.

[RETRACE] Highest-ranked verification assay

Rank: 1
Name:
Trace visual provenance of the Hurricane Harvey shark image
Sub-task:
image / cross-modal
Target:
Whether the shark shown in the flooded-road photograph was genuinely photographed in Houston during Hurricane Harvey.

Preferred sources:
- Original photographer or photographic archive
- Earlier versions of the same shark image
- Independent reporting documenting image provenance

Expected resolution:
Identify the original source of the shark and determine whether the viral Houston photograph is genuine or a composite.

Reasoning:
If the shark originates from an independently documented photograph created before Hurricane Harvey, then the viral image's claimed Houston provenance can be directly tested.

[RETRACE] Generating candidate investigation tasks.

[RETRACE] Ranking investigation tasks with pairwise comparison + BTL aggregation.

[RETRACE] Top-ranked investigation tasks

1. Recover the original source of the shark
BTL strength score: 0.81
Sub-task: image
Target: Determine whether the same shark occurs in an earlier photograph.

2. Trace reuse of the shark in Hurricane Harvey imagery
BTL strength score: 0.67
Sub-task: image+text
Target: Determine whether credible sources document this shark being composited into the Houston flood image.

3. Verify the Houston flood context independently
BTL strength score: 0.42
Sub-task: text
Target: Determine whether Hurricane Harvey genuinely caused severe flooding in Houston.

Figure 5: Exemplary RETRACE fact-check report illustrating the dual-stage investigation and a correct final prediction, page 1 of 4.

========== STAGE I LATS INVESTIGATION ==========

Iteration 1
Thought 1: The decisive uncertainty is the shark's provenance, so I should first identify an earlier source photograph.

Action 1: Google["great white shark kayak Thomas Peschak original photograph"]

Observation 1: Thomas P. Peschak documents an authentic photograph of a great white shark following a yellow kayak off South Africa. His account identifies the photograph as an original film image and describes its later reuse in misleading contexts.

Iteration 2
Thought 2: The source shark is identified; I need evidence connecting that photograph specifically to the Hurricane Harvey composite.

Action 2: Google["Thomas Peschak shark Hurricane Harvey Houston flooded street"]

Observation 2: National Geographic reports that Peschak's shark was repeatedly inserted into fabricated disaster photographs and specifically resurfaced in Houston flood imagery during Hurricane Harvey in 2017.

Iteration 3
Thought 3: The manipulated image is established, but the real flood event should be separated from the false visual provenance.

Action 3: Google["Hurricane Harvey Houston flooding August 2017 official report"]

Observation 3: Authoritative records document severe Hurricane Harvey flooding in the Houston region. Therefore, the event context is genuine, while the shark's association with it remains contradicted by provenance evidence.

========== STAGE II: EVIDENCE-GAP DIAGNOSIS==========

Factual Requirement Extraction

[REQUIREMENT PIPELINE] Extracting independently verifiable requirements.

Requirement (p1)
Proposition: Hurricane Harvey caused severe flooding in Houston in August 2017.

Verification type: event / time / location

Modality: T

Requirement (p2)
Proposition: The displayed shark photograph was captured in Houston during Hurricane Harvey.

Verification type: image provenance

Modality: TI

Requirement (p3)
Proposition: The shark is an authentic part of the flooded-road photograph rather than a composited element.

Verification type: manipulation / provenance

Modality: I

Requirement (p4)
Proposition: The shark does not originate from an earlier unrelated photograph.

Verification type: source identity

Modality: I

Figure 6: Exemplary RETRACE fact-check report illustrating the dual-stage investigation and a correct final prediction, page 2 of 4.

========== MULTIVIEW EMOTION CHARACTERIZATION ==========

[RETRACE] Analysing textual view.

[RETRACE] Analysing visual view.

[RETRACE] Analysing joint image-text view.

[RETRACE] Cross-modal discrepancy

========== REQUIREMENT-LEVEL EVIDENCE COVERAGE ==========

Requirement p1
Coverage c1: 1.00
Direction d1: SUPPORT
Supporting evidence: E1

Assessment: The occurrence and location of severe Hurricane Harvey flooding are decisively established.
Requirement p2
Coverage c2: 0.75
Direction d2: CONTRADICT
Supporting evidence: E2, E3

Assessment: Strong evidence indicates that the image was not captured as claimed, but direct provenance from the original photographer should be retained as decisive evidence.
Requirement p3
Coverage c3: 0.75
Direction d3: CONTRADICT
Supporting evidence: E2, E3

Assessment: Existing evidence strongly indicates compositing, but the exact source relationship should be resolved through primary-source provenance.
Requirement p4
Coverage c4: 1.00
Direction d4: CONTRADICT
Supporting evidence: E2

========== STRUCTURED GAP DIAGNOSIS ==========

[RETRACE] The unresolved requirements are

p2: Was this image genuinely captured in Houston during Harvey?

p3: Is the shark physically part of the original flood photograph? Their incomplete coverage makes them candidates for repair.

[RETRACE] Highest-priority gap
Name: Establish decisive primary-source provenance for the shark element

Target: Confirm from the original photographer/source that the shark predates Hurricane Harvey and was reused in a composite.

Why selected: The claim's strongest persuasive element is visual. Existing evidence already contradicts the image provenance, but direct creator evidence would decisively close the remaining gap.

Figure 7: Exemplary RETRACE fact-check report illustrating the dual-stage investigation and a correct final prediction, page 3 of 4.

========== **STAGE II LATS REPAIR SEARCH** ==========

Thought 1: Direct photographer provenance can decisively determine whether the shark predates the claimed Houston event.

Action 1: Google[site:thomaspeschak.com great white shark kayak flood photograph provenance]

Observation 1: Peschak explains that his original great-white-shark photograph was created in South Africa and later reused in manipulated flood scenes. His site specifically describes the shark being superimposed into an unrelated flooded street image.

========== **FINAL EVIDENCE-GROUNDED VERIFIER** ==========

Verifier input

Claim: A shark was photographed swimming on a flooded highway in Houston during Hurricane Harvey.

Evidence: E1, E2, E3, E4, E5

Final explanation

The claim combines a genuine disaster with manipulated visual evidence. Hurricane Harvey caused severe flooding in Houston, but the shark shown in the viral image was not photographed there. The animal originates from Thomas P. Peschak's earlier photograph of a great white shark following a kayak in South Africa. Peschak's provenance account and independent National Geographic reporting document the later reuse of this shark in fabricated disaster imagery, including during Hurricane Harvey. The claimed visual provenance is therefore contradicted by the recovered evidence.

Judge label: REFUTED

Figure 8: Exemplary RETRACE fact-check report illustrating the dual-stage investigation and a correct final prediction, page 4 of 4.



News Text: Breaking: Large explosion near the Pentagon complex in Washington, D.C.
Image: Viral image showing a large plume of dark smoke beside a government-looking building.
GT: REFUTED
PRED: REFUTED

Claim Report

The multimodal post claims that a large explosion occurred near the Pentagon and presents an image of dark smoke as visual evidence. The claim circulated online on May 22, 2023, but authorities responsible for the Pentagon area reported that no explosion or incident had occurred. Arlington County Fire and EMS and the Pentagon Force Protection Agency publicly stated that there was no explosion at or near the Pentagon and no immediate danger to the public. Subsequent Department of Defense remarks referred to the circulating material as bogus imagery of an alleged Pentagon explosion. Independent reporting further found that the building in the viral image did not correspond convincingly to the Pentagon and that visual anomalies were consistent with an apparently AI-generated image. Thus, both the claimed event and the purported visual evidence are contradicted, supporting the final verdict REFUTED.

References:
[1] Arlington County Fire and EMS / Pentagon Force Protection Agency statement, reported by NPR: no explosion or incident was occurring at or near the Pentagon.
[2] U.S. Department of Defense press briefing referring to the circulating Pentagon-explosion images as “bogus images.”
[3] CNN reporting on structural inconsistencies in the viral image and expert assessment that it was likely AI-generated.
[4] Washington Post reconstruction of the incident, including Arlington emergency-system checks showing no corresponding fire, medical, or emergency calls.

========== RETRACE DECISION TRACE ==========
========== STAGE I: VERIFICATION STRATEGY DISCOVERY ==========

[ASSAY PIPELINE] Starting investigation.

[STRICT ROBIN] Generating candidate verification assays.

[STRICT ROBIN] Ranking assays with pairwise comparison + BTL aggregation.

[RETRACE] Highest-ranked verification assay

Rank: 1
BTL strength score: 1.31

Name: Verify whether a Pentagon-area explosion actually occurred

Sub-task: text / event verification

Target: Whether a large explosion occurred at or near the Pentagon at the time claimed by the viral post.

Preferred sources:
- Arlington County Fire and EMS
- Pentagon Force Protection Agency
- U.S. Department of Defense
- Contemporaneous local reporting

Expected resolution: Determine whether emergency and security agencies recorded an explosion or explicitly denied that one occurred.

Reasoning: Event verification is more decisive than inspecting the image alone: if the responsible authorities confirm that no explosion occurred, the core factual claim is directly contradicted.

[RETRACE] Generating candidate investigation tasks.

[RETRACE] Ranking investigation tasks with pairwise comparison + BTL aggregation.

[RETRACE] Top-ranked investigation tasks

1. Check authoritative emergency-service confirmation
BTL strength score: 0.83
Sub-task: text
Target: Determine whether Arlington emergency services recorded an explosion near the Pentagon.

2. Check Pentagon / Department of Defense confirmation
BTL strength score: 0.71
Sub-task: text
Target: Determine whether Pentagon security or DoD acknowledged an actual incident.

3. Investigate image authenticity
BTL strength score: 0.54
Sub-task: image
Target: Determine whether the building, fence, smoke, and architectural geometry are consistent with a genuine Pentagon photograph.

4. Trace the original reporting account
BTL strength score: 0.39
Sub-task: text / source
Target: Determine whether the account circulating the initial claim was genuinely affiliated with a recognized news organization.

Figure 9: Second exemplary RETRACE case study illustrating a correct evidence-grounded prediction, page 1 of 4

========== **STAGE I LATS INVESTIGATION** ==========

Iteration 1
Thought 1: A real explosion near the Pentagon should produce immediate emergency-service records, so I need authoritative incident confirmation..

Action 1: Google["Pentagon explosion Arlington Fire EMS May 22 2023 official"]

Observation 1: Arlington County Fire and EMS and the Pentagon Force Protection Agency stated that there was no explosion or incident taking place at or near the Pentagon reservation and no immediate danger to the public.

Iteration 2
Thought 2: The emergency-service denial directly contradicts the event, but Pentagon-level confirmation would independently strengthen the finding.

Action 2: Google["Pentagon May 22 2023 explosion Department Defense official statement"]

Observation 2: In a subsequent official Department of Defense briefing, Pentagon Press Secretary Brig. Gen. Pat Ryder referred to the circulating material as bogus images of alleged explosions at the Pentagon.

Iteration 3
Thought 3: The physical event is refuted; I should now determine whether the viral image itself provides any genuine Pentagon evidence.

Action 3: Google["Pentagon explosion viral image AI structural errors building May 2023"]

Observation 3: Independent analysis reported that the building shown did not closely resemble the Pentagon and contained structural irregularities associated with synthetic imagery. Digital-forensics expert Hany Farid described architectural and fence anomalies as typical signs of AI-generated imagery.

========== **STAGE II: EVIDENCE-GAP DIAGNOSIS** ==========

Factual Requirement Extraction

[REQUIREMENT PIPELINE] Extracting independently verifiable requirements.

Requirement (p1)
Proposition: A large explosion occurred at or near the Pentagon on May 22, 2023.

Verification type: event / official confirmation

Modality: T

Requirement (p2)
Proposition: The viral image depicts the Pentagon or its immediate surroundings.

Verification type: location / visual identity

Modality: I

Requirement (p3)
Proposition: The smoke and explosion shown in the image are authentic photographic evidence of the claimed event.

Verification type: image authenticity

Modality: I

Requirement (p4)
Proposition: The breaking-news report originated from a legitimate authoritative news source.

Verification type: source identity / provenance

Modality: TI

Figure 10: Second exemplary RETRACE case study illustrating a correct evidence-grounded prediction, page 2 of 4

========== MULTIVIEW EMOTION CHARACTERIZATION ==========

[RETRACE] Analysing textual view.

[RETRACE] Analysing visual view.

[RETRACE] Analysing joint image-text view.

[RETRACE] Cross-modal discrepancy

========== REQUIREMENT-LEVEL EVIDENCE COVERAGE ==========

Requirement p1
Coverage c1: 1.00
Direction d1: CONTRADICT
Supporting evidence: E1, E2, E3

Assessment: Multiple responsible authorities explicitly state that no explosion occurred.
Requirement p2
Coverage c2: 0.50
Direction d2: CONTRADICT
Supporting evidence: E4

Assessment: Visual inconsistencies suggest that the scene does not depict the Pentagon, but additional landmark comparison would improve resolution.
Requirement p3
Coverage c3: 0.75
Direction d3: CONTRADICT
Supporting evidence: E3, E4

Assessment: The image is strongly inconsistent with an authentic photograph of the claimed event, although its exact generation provenance is not known.
Requirement p4
Coverage c4: 0.50
Direction d4: CONTRADICT
Supporting evidence: partial source-tracing evidence

========== STRUCTURED GAP DIAGNOSIS ==========

[RETRACE] The unresolved requirements are

p2: Does the pictured structure actually correspond to the Pentagon?

p4: Did the breaking-news report originate from a genuine news organization?

Figure 11: Second exemplary RETRACE case study illustrating a correct evidence-grounded prediction, page 3 of 4

========== STAGE II TARGETED REPAIR PLANNING==========

Repair Action 1
Name: Pentagon landmark-consistency verification

Target: Determine whether the government-looking building in the viral image matches the real Pentagon.

Repair Action 2
Name: Original-source attribution

Target: Determine whether the account that first amplified the breaking-news claim was genuinely affiliated with Bloomberg or another recognized news organization.

========== STAGE II LATS REPAIR SEARCH ==========

Repair Iteration 1
Thought 1: The event itself is already refuted, but landmark verification can determine whether the image even visually represents the Pentagon.

Action 1: Google["viral Pentagon explosion image building does not resemble Pentagon"]

Observation 1: Reporting and visual-forensics analysis found that the building in the viral image did not closely match the Pentagon. Structural inconsistencies in the building and fence were among the features cited as indications of synthetic generation.

Repair Iteration 2
Thought 2: The event and image are contradicted; I need to establish whether apparent newsroom authority helped the false report spread.

Action 2: Google["Bloomberg Feed fake account Pentagon explosion May 22 2023"]

Observation 2: Contemporaneous reporting found that a verified-looking account calling itself "Bloomberg Feed" circulated the claim but was not affiliated with Bloomberg News. The false image was subsequently amplified by other accounts and media outlets. The Washington Post

========== FINAL EVIDENCE-GROUNDED VERIFIER ==========

Verifier input

Claim: A large explosion occurred near the Pentagon, as shown in the attached image.

Evidence: E1, E2, E3, E4, E5

Final explanation

The claim is contradicted by both event-level and visual-provenance evidence. Arlington County Fire and EMS and Pentagon Force Protection stated that no explosion or incident occurred at or near the Pentagon. The U.S. Department of Defense subsequently described the circulating explosion imagery as bogus. Independent visual analysis also found that the depicted structure did not convincingly match the Pentagon and contained anomalies consistent with synthetic imagery, while source tracing showed that one prominent account amplifying the story was not actually affiliated with Bloomberg. The post therefore describes an event that did not occur and uses non-authentic visual evidence to support it.

Judge label: REFUTED

Figure 12: Second exemplary RETRACE case study illustrating a correct evidence-grounded prediction, page 4 of 4