REVEAL: Retrieval-Enhanced Verification for Multimodal Fact-Checking

Amina Tariq and Yova Kementchedjhieva
Mohamed bin Zayed University of Artificial Intelligence

Abstract

Multimodal misinformation combines images and text to amplify false narratives, yet most fact-checking research addresses only textual claims. The AVerImaTeC shared task introduces real-world image-text claims requiring sophisticated evidence retrieval. We present **REVEAL** (**R**etrieval-**E**nhanced **V**erification with **E**vidence **A**ccumulation **L**oop), a system designed to overcome the "semantic gap," defined as the disconnect between the neutral phrasing of claims and the adversarial vocabulary of debunking evidence. Unlike static baselines, REVEAL breaks down the verification task into an iterative context loop, integrating sparse and dense retrieval signals to aggressively target refuting evidence. We achieve a Verdict Accuracy of 23.6% and an Evidence Recall of 27.7% on the test set. Our results outperform the official baseline across all metrics, validating our hybrid retrieval strategy for complex multimodal verification.

1 Introduction

The rapid proliferation of multimodal misinformation poses a critical threat to the information ecosystem. Recent studies indicate that approximately 80% of online claims combine text and images to construct narratives (Dufour et al., 2024), yet automated fact-checking (AFC) research has predominantly focused on textual verification. To address this gap, the AVerImaTeC shared task (Cao et al., 2025) introduces a benchmark of 1,297 real-world image-text claims, requiring systems to retrieve multimodal evidence and predict verdicts grounded in a static knowledge store.

The task's official baseline relies on sparse retrieval methods like BM25 (Robertson and Zaragoza, 2009) for text and CLIP (Radford et al., 2021) for image alignment. While these methods provide a robust foundation for general similarity search, we observe that they lack the specificity required for adversarial verification. This results in a lexical disparity: the neutral language of a claim rarely matches the charged vocabulary of debunking evidence (e.g., "hoax" or "false"). Furthermore, generic vision encoders often fail to distinguish between visually similar but contextually distinct entities, such as specific highways or buildings.

We propose REVEAL, a system designed to bridge this divide by augmenting robust sparse signals with advanced semantic reasoning. Unlike static baselines, REVEAL employs a dynamic pipeline that combines Hypothetical Document Embeddings (HyDE) (Gao et al., 2023) and a Weighted Multi-Query strategy to aggressively target debunking evidence. We further refine these results using Dense Reranking to favor authoritative domains, while leveraging SigLIP2 (Tschannen et al., 2025) to ensure precise visual grounding.

REVEAL achieves 23.6% Verdict Accuracy and 27.7% Evidence Recall on the AVerImaTeC test split. We show that while BM25 remains a powerful signal, significantly outperforming the baseline requires augmenting it with diversity-driven query expansion and weighted fusion strategies.

2 System Architecture

2.1 The Baseline Framework

Our system builds upon the official AVerImaTeC baseline (Cao et al., 2025), a modular agentic framework for multimodal claim verification that operates on a *decompose-and-verify* paradigm. In each iteration, a Question Generator produces a verification question which a Planner LLM maps to: (A) Reverse Image Search (RIS), (B) Visual Question Answering (VQA), (C) Text Search, or (D) Image Search. The retrieved data is aggregated into a running evidence history, finally passed to a Verifier module for a verdict.

While this architecture provides a logical skeleton, the baseline implementation relies on *naive*

Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 108–113

execution with two critical limitations identified in (Cao et al., 2025):

1. **Stateless & Biased Retrieval:** The dynamic strategy failed to outperform parallel generation, and the tool selector showed a 30% bias toward VQA, often neglecting external search.

2. **Generic Encoders & Retrieval Gaps:** Reliance on BM25 ranking ignores visual content. Consequently, 13% of claims retrieved no context via RIS, a hard ceiling that standard tools cannot bypass.

REVEAL addresses these by replacing this naive toolset with a **history-aware** retrieval pipeline, where evidence retrieved in earlier turns is summarized and explicitly reused to condition subsequent question generation, retrieval, and verification steps (Figure 1).

2.2 Addressing Vocabulary Mismatch with HyDE

A primary challenge in automated fact-checking is the misalignment between a verifier's questions (e.g., "Is this image authentic?") and the actual evidence (e.g., "Deepfake generated by Midjourney"). To address this, we leverage HyDE (Gao et al., 2023).

Rather than searching directly with the raw question, we prompt an LLM to hallucinate a theoretical "perfect" evidence snippet (Appendix A.1). This generates a semantic target that captures the likely patterns of relevant fact-checks. These generated passages, along with potential visual descriptions, are added to our search pool to improve recall for non-keyword-based matches.

2.3 Weighted Multi-Query Retrieval

To maximize the retrieval of conflicting evidence, we move beyond simple query matching. We implement a Weighted Multi-Query strategy that executes four distinct query types in parallel, each assigned a specific importance weight (w_q):

- **Debunking Heuristics** ($w = 3.0$)**:** We explicitly construct adversarial queries by appending terms like "*false*", "*hoax*", and "*fact check*" to the claim. These are weighted highest to prioritize documents that directly refute the claim.

- **Primary Question** ($w = 2.0$)**:** The specific verification question generated by the LLM.

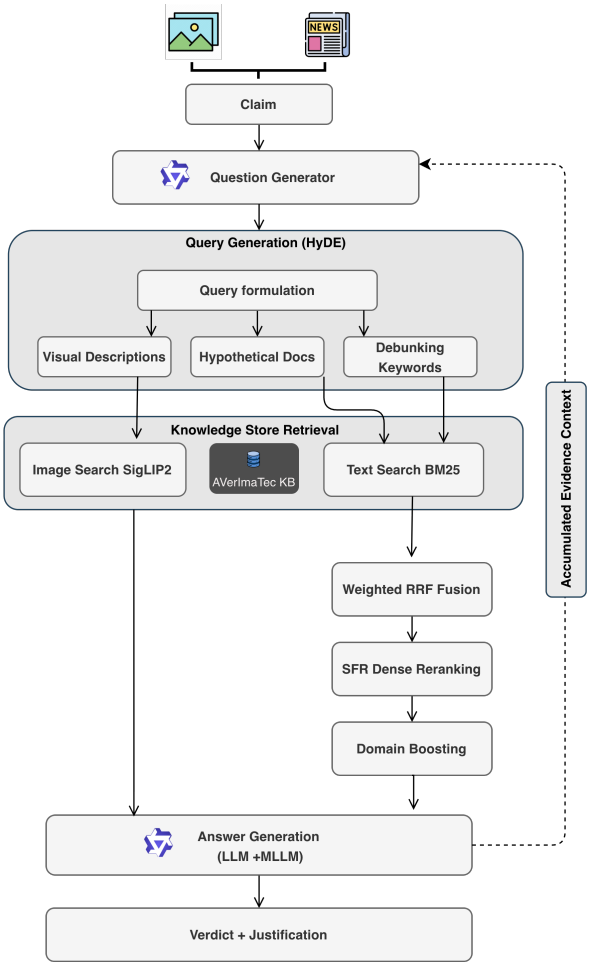

Figure 1: **System Architecture of REVEAL.** The pipeline features (1) HyDE-driven query expansion to bridge the semantic gap, (2) a Hybrid Retrieval module combining SigLIP2, weighted BM25, and dense reranking, and (3) an Evidence Accumulation Loop where retrieved context feeds back into the question generator.

- **Claim Context** ($w = 1.5$)**:** The raw text of the original claim to ensure topical relevance.

- **HyDE Augmentation** ($w = 1.0$)**:** Hypothetical documents generated in the prior step.

Results are aggregated using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). Unlike standard RRF, we introduce the weight term w_q to bias the ranking toward our debunking heuristics:

$$Score(d) = \sum_{q \in Q} w_q \times \frac{1}{k + r_q(d)} \quad (1)$$

where $k = 60$ and $r_q(d)$ is the rank of document d in the result list for query q. This weighting scheme ensures that a document matching a "debunking" query bubbles to the top, even if it appears lower in the raw BM25 rankings. These weights were

empirically tuned on the development set to prioritize debunking signals and optimize the retrieval of contradictory evidence for the verification task.

2.4 Dense Reranking & Domain Boosting

While sparse retrieval ensures high recall, it often returns noisy results. We refine the top 500 candidates using SFR-Embedding-2_R (Rui Meng*, 2024), computing cosine similarity to filter out irrelevant matches. We then select the top 15 reranked documents as the final textual evidence set passed to the verifier.

To further emulate human verification behavior, we apply a heuristic Domain Boost. We maintain a whitelist of authoritative fact-checking organizations (e.g., *Snopes, PolitiFact, Reuters*). Evidence retrieved from these domains receives an empirically determined 1.5× similarity boost, ensuring that high-trust sources override random blog posts or social media commentary in the final context window.

2.5 Visual Grounding with SigLIP2

Standard baselines typically rely on CLIP for image retrieval, which often struggles with fine-grained entity recognition. We replace this component with SigLIP2-Large (patch16-384), selected for its superior zero-shot performance in multilingual and noisy web environments.

For image-related questions, both the textual query and HyDE-generated visual descriptions are encoded into the SigLIP embedding space, enabling retrieval based on semantic similarity (e.g., identifying images of a specific event) rather than just visual overlap. As shown in our ablation study (Table 2), upgrading the visual encoder alone yields only marginal gains, suggesting that SigLIP2's primary value is in providing a more reliable grounding space for HyDE-driven semantic expansions, which remain the dominant contributor to retrieval performance.

2.6 Iterative Verification Loop

In the final stage, the retrieved multimodal evidence is processed by Qwen2.5-VL (Bai et al., 2025). To ensure the model builds upon prior findings, REVEAL implements an Evidence Accumulation Loop in which the verifier LLM produces a concise textual summary of the evidence from previous turns, and this summary is included in the context for subsequent question generation (Appendix A.2). This iterative process enables the system to adapt its search strategy, either terminating early upon finding definitive debunking evidence or pivoting to alternative lines of inquiry when results remain inconclusive.

3 Experimental Results

Experimental Setup. We evaluate REVEAL on the AVerImaTeC dataset (Cao et al., 2025), utilizing the development set (152 claims) for ablation studies and the official test set (352 claims) for leaderboard ranking. The dataset contains significant label imbalance, with 92.8% of development claims being *Refuted*. Our backbone is Qwen2.5-VL-7B-Instruct[1], deployed in BF16 precision across three NVIDIA RTX 5000 Ada GPUs. We use a generation temperature of 0.3 and nucleus sampling with $top\_p = 0.9$.

We compare REVEAL with the Official Challenge Baseline (architecture detailed in §2.1). To isolate the impact of our contributions, we evaluate the cumulative addition of our components: the *Local Baseline* replicates the unenhanced pipeline using BM25 and CLIP; + *SigLIP2* upgrades the visual encoder[2]; + *HyDE* introduces semantic expansion; + *Weighted RRF* prioritizes debunking terms; and the *Full System* incorporates dense reranking via SFR-Embedding-2[3].

Challenge Results (Test Set). Table 1 presents the top entries from the official challenge leaderboard. REVEAL achieves a Verdict Accuracy of 23.6% and an Evidence Score of 0.277. Notably, our system significantly outperforms the Official Challenge Baseline, achieving a 107% relative improvement in Verdict Accuracy (23.6% vs 11.4%) and a 62% improvement in Evidence Retrieval (0.277 vs 0.171). This validates that while the baseline architecture is sound, its standard retrieval tools are insufficient for complex multimodal verification without the state-aware enhancements introduced by REVEAL.

Ablation Analysis (Dev Set). The cumulative impact of our design choices is detailed in Table 2. The introduction of HyDE provided the single largest jump in Evidence Score (0.245 → 0.279), confirming that hypothetical document expansion

[1] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

[2] https://huggingface.co/google/siglip2-large-patch16-384

[3] https://huggingface.co/Salesforce/SFR-Embedding-2_R

System	Q-Score	Evid. Score	Verd. Acc	Just. Score
HUMANE (Rank 1)	0.890	0.536	0.546	0.556
ADA-AGGR	0.370	0.463	0.537	0.433
AIC CTU	0.807	0.325	0.347	0.304
REVEAL (Ours)	**0.632**	**0.277**	**0.236**	**0.135**
Challenge Baseline	0.555	0.171	0.114	0.132

Table 1: Official results on the AVerImaTeC Test Set (352 claims). REVEAL (Rank 6) more than doubles the baseline verdict accuracy.

effectively bridges the vocabulary mismatch. In contrast, replacing CLIP with SigLIP2 alone yields only a marginal improvement in Evidence Score, suggesting that visual encoder upgrades primarily play a supporting role, while semantic expansion via HyDE is the dominant contributor. While the application of Weighted RRF resulted in a slight reduction in raw evidence recall ($0.279 \rightarrow 0.260$) due to the suppression of lower-ranked documents, it importantly improved Verdict Accuracy ($0.250 \rightarrow 0.257$). This suggests that the consensus strategy successfully filtered out distracting noise. Finally, the Full System incorporating the domain-adapted SFR reranker achieved the highest performance across all metrics (30.3% Accuracy), validating that dense reranking is essential for identifying semantically distinct but factually critical evidence.

Configuration	Q-Score	Evid.	Verd.	Just.
Baseline (BM25 + CLIP)	0.602	0.244	0.211	0.093
+ SigLIP2	0.639	0.245	0.217	0.115
+ HyDE	0.641	0.279	0.250	0.117
+ Weighted RRF	0.637	0.260	0.257	0.124
+ SFR2 Reranker (Full)	**0.652**	**0.292**	**0.303**	**0.132**

Table 2: Ablation study on the development set. The stepwise integration of SigLIP2, HyDE, RRF, and SFR2 yields consistent improvements, with the Full System achieving the highest accuracy.

Qualitative Analysis. Qualitative inspection of failure cases reveals distinct error mechanisms linked to our retrieval strategy. The most common failure mode, *Safety Bias* (Refuted $\rightarrow$ NEI), occurs when retrieval returns related context but lacks a definitive "debunking" link. For example, regarding the claim *"Images show the renovation of the Cuttack hospital just before PM Modi's arrival,"* REVEAL correctly retrieved reports of Modi's visit but failed to find specific evidence linking the renovation photos to that event. Lacking a "smoking gun," the verifier conservatively abstained.

Conversely, we observed *Aggressive Debunking* (Supported $\rightarrow$ Refuted), where our weighted queries ($w = 3.0$) caused the model to over-

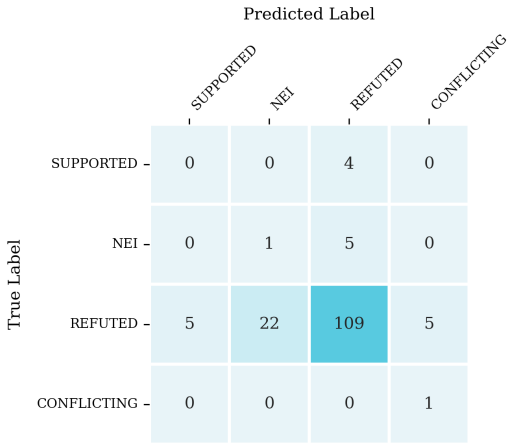

Figure 2: **Confusion Matrix on the Development Set.** High raw accuracy (73%) reflects correct identification of the majority class (*Refuted*), yet frequent *NEI* errors reveal retrieval gaps.

interpret minor discrepancies. For the claim *"Finland's Kummakivi Rock has been stood on top of the rock below for 11,000 years,"* the system retrieved evidence stating the rock is 11,500 years old. The verifier predicted *Refuted*, likely penalizing the slight numerical difference or reacting to "hoax" terms retrieved in the search context. Finally, the *Visual Semantic Gap* remains challenging. For *"Image of Jammu National Highway 44,"* the system mistook generic mountain roads for the specific location, resulting in a *False Support*.

4 Discussion & Conclusion

Our work demonstrates that standard open-source multimodal baselines can be significantly elevated through targeted retrieval enhancements. By replacing generic tools with domain-specific components, such as SigLIP2 for visual grounding and Weighted RRF for debunking, REVEAL consistently outperforms the baseline. This confirms that the primary bottleneck in automated fact-checking is often not the multimodal large language model (MLLM)'s reasoning capability, but the specificity of the evidence fed into its context window.

However, the iterative nature of this process introduces a notable computational trade-off. Inference logs show that while visual retrieval is efficient ($\approx$4s), text retrieval averages 35–45s per turn (RTX 5000) due to SFR reranking. With sequential processing, complex claims take $\approx$3 minutes per claim. While this limits real-time scalability, such depth is requisite for sophisticated misinformation. Consequently, future work should focus

on: (1) distilling dense rerankers to reduce this bottleneck, and (2) parallelizing the QA loop to improve throughput without sacrificing reasoning depth.

Finally, we highlight the friction between strict evaluation metrics and latent reasoning. The protocol enforces a dual requirement: a valid prediction requires both a correct verdict *and* evidence surpassing the threshold λ. Consequently, instances where the model correctly debunks a claim using partial or inferred context are treated as failures. As shown in the confusion matrix (Figure 2), this rigid scoring often obscures the model's practical ability to identify misinformation.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. Averimatec: A dataset for automatic verification of image-text claims with evidence from the web. *arXiv preprint arXiv:2505.17978*.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *CoRR*, abs/2405.11697.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng*, Ye Liu*. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

A Prompt Templates

A.1 HyDE: Multimodal Evidence Generation

Our implementation of HyDE (Gao et al., 2023) uses three distinct prompts to generate diverse retrieval targets.

HyDE 1: Hypothetical Fact-Check

System: You are an expert OSINT investigator analysing multimodal misinformation. Write short, neutral fact-checking style paragraphs.

User: Claim: [CLAIM_TEXT]
Detected Text (OCR): [OCR_HINT]
You are drafting a SHORT, NEUTRAL FACT-CHECKING NOTE. Write 2–3 sentences (max 80 words) that:

- Describe what the images show objectively.
- Mention key people, places, dates, or text.
- Indicate likely real-world context.
- If suspicious, hint at actual context.

Do NOT give instructions like "search for".
Output ONLY the final paragraph.

Figure 3: Prompt for fact-check generation.

HyDE 2: Visual Search Queries

System: You are a reverse image search specialist. Output compact keyword phrases to find the TRUE origin of images.

User: Claim: [CLAIM_TEXT]
OCR Text: [OCR_HINT]
Task: Generate 5–7 optimized keyword phrases for reverse image search to verify the claim.

- Identify potential misinformation indicators (e.g., temporal, spatial, or identity mismatches).
- Focus on specific visual entities, text, or landmarks visible in the image.
- Prioritize phrases that target the *actual* context rather than the claimed context.

Format: Single line of comma-separated phrases (3–6 words each).

Figure 4: Prompt for visual query generation.

HyDE 3: Hypothetical Captions

System: You are a neutral image captioning assistant for fact-checkers.

User: Claim: [CLAIM_TEXT]
OCR hint: [OCR_HINT]
Task: Generate three objective captions describing what *true* evidence images for this claim might show.

- Focus on objective visual descriptors (who/what/where).
- Hypothesize alternative contexts (e.g., "photo from a different event").
- Maintain a strictly neutral tone without asserting a verdict.

Format: One caption per line (max 30 words each).

Figure 5: Prompt for caption generation.

A.2 Question Generation

The Question Generator formulates the next step in the inquiry chain based on the current evidence state.

Question Generation (Iterative Loop)

System: You are a professional fact-checker verifying an image-text claim. Your goal is to uncover the truth by asking targeted follow-up questions.

User: Claim: [CLAIM_TEXT]
Evidence So Far: [EVIDENCE_HISTORY]

Task: Formulate the next logical question to advance the verification process.

- Analyze the provided *Evidence So Far* to avoid redundant inquiries.
- Determine if the next step should verify visual details (e.g., landmarks, text in image) or textual assertions.
- If previous evidence is inconclusive, pivot to a new line of inquiry.

Format: Output exactly one question prefixed with its type: ***Image-related:** [Question]* OR ***Text-related:** [Question]*

Figure 6: Prompt for iterative question generation.