Selective Multimodal Retrieval for Automated Verification of Image–Text Claims

Yoana Tsoneva[1], Paul-Conrad Feig [1], Jiaao Li[1], Veronika Solopova[1,2], Neda Foroutan[1], Arthur Hilbert[1,2], Vera Schmitt[1,2,3,4]
[1] Technische Universität Berlin
[2] German Research Center for Artificial Intelligence (DFKI)
[3] BIFOLD – Berlin Institute for the Foundations of Learning and Data
[4] Centre for European Research in Trusted AI (CERTAIN)
Correspondence: yoana.tsoneva@tu-berlin.de

Abstract

This paper presents an efficiency-aware pipeline for automated fact-checking of real-world image–text claims that treats multimodality as a controllable design variable rather than a property that must be uniformly propagated through every stage of the system. The approach decomposes claims into verification questions, assigns each to text- or image-related types, and applies modality-aware retrieval strategies, while ultimately relying on text-only evidence for verdict prediction and justification generation. Evaluated on the AVerImaTeC dataset within the FEVER-9 shared task, the system achieves competitive scores and ranks fourth overall, outperforming the official baseline on Evidence Score (0.2703 vs. 0.1707), Verdict Accuracy (0.2557 vs. 0.1136), and Justification Score (0.198 vs. 0.1322). These results demonstrate that strong performance on multimodal fact-checking can be achieved by selectively controlling where visual information influences retrieval and reasoning, rather than performing full multimodal fusion at every stage of the pipeline.

1 Introduction

Automated fact-checking has seen substantial progress in recent years, largely driven by advances in retrieval-based pipelines and large language models for textual claim verification. Benchmarks such as FEVER (Thorne et al., 2018) and AVeriTeC (Schlichtkrull et al., 2024) have established standardized evaluation settings in which systems retrieve textual evidence, predict claim veracity, and generate justifications. While these approaches perform well for text-only claims, they are increasingly challenged by real-world misinformation that couples textual statements with visual content. Image-text claims introduce additional sources of ambiguity and error. Visual material may be authentic but miscaptioned, taken from a wrong context, or framed to manipulate emotionally or support a false narrative. Recent analyses of fact-checking datasets show that a non-trivial portion of claims require reasoning over visual context to reach a correct verdict, motivating the development of multimodal benchmarks such as AVerImaTeC (Cao et al., 2025).

Existing multimodal fact-checking systems commonly address this challenge by integrating vision-language models and multimodal retrieval components throughout the pipeline. While effective, these designs often lead to high computational cost and introduce additional sources of instability, e.g., through sensitivity to prompt formulation or implicit modality fusion.

In this work, we investigate a pragmatic alternative. Instead of treating multimodality as a property that must be propagated through all stages of the system, we explicitly control where and how visual information is used. Our approach distinguishes between text-related and image-related verification questions and applies modality-aware retrieval strategies accordingly, while aggregating primarily textual evidence for final verdict prediction and justification. This design aims to retain the benefits of multimodal reasoning where it is most informative, without incurring the full complexity of fully multimodal evidence fusion.

We evaluate our approach on the AVerImaTeC dataset and show that competitive performance can be achieved using this selective multimodal strategy. Our results highlight that strong multimodal fact-checking does not necessarily require multimodal processing at every stage, and that pipeline-level control over modality usage offers an effective and efficient design choice. Our system achieved fourth place in the FEVER-9 shared task[1]. Specifically, we achieved a 58.3% relative improvement in Evidence Score (0.2703 vs. 0.1707) and more

[1] https://fever.ai/workshop.html

Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 127–135

than doubled the conditional Verdict Accuracy on the test set (0.2557 vs. 0.1136). Furthermore, our system reached a Justification Score of 0.198, representing a 50% relative improvement over the baseline score of 0.1322.

The contributions of this paper are as follows[2]:

- We propose a selective multimodal fact-checking framework that explicitly controls where and how visual information is incorporated within the verification pipeline.
- We introduce an efficiency-aware design that applies modality-specific retrieval strategies while relying primarily on textual evidence for verdict prediction and justification generation.
- Through experiments on the AVerImaTeC benchmark, we demonstrate that this selective use of visual information improves evidence quality, verdict accuracy, and justification coherence compared to the FEVER-9 baseline.

2 Related Work

Automated fact-checking (AFC) has traditionally focused on textual claims, with seminal benchmarks such as FEVER (Thorne et al., 2018) and subsequent shared tasks like AVeriTeC (Schlichtkrull et al., 2024) establishing a standardized pipeline of claim verification, evidence retrieval, verdict prediction, and justification generation. Recent advances in large language models (LLMs) have further improved performance on text-based fact-checking, including zero-shot and instruction-tuned approaches, such as in Upravitelev et al. (2025), which inspired our work. However, several studies demonstrate that purely textual fact-checking systems struggle with real-world claims that rely on visual context or cross-modal inconsistencies. Large-scale analyses show that a substantial portion of fact-checkable claims require multimodal reasoning, particularly involving images, and cannot be reliably verified using text alone (Akhtar et al., 2023; Cao et al., 2025). Empirical studies further indicate that images increase the perceived credibility of false claims, amplifying misinformation beyond what text-only models can capture (Hameleers et al., 2020). Recent benchmark results confirm that multimodal retrieval and reasoning significantly improve verification accuracy for image-text claims compared to text-only pipelines (Cekinel et al., 2025). This has led to the development of multimodal fact-checking benchmarks and shared tasks, such as new AVerImaTeC (Cao et al., 2025).

Recent high-performing systems rely on modular, pipeline-based architectures. In parallel, research on multimodal representations examines how visual and textual information should be combined for fact-checking. Akhtar et al. (Akhtar et al., 2023) formalize multimodal AFC as a staged process and demonstrate that naive modality fusion is frequently inferior to structured cross-modal reasoning and retrieval-based approaches. Recent studies explore the role of vision-language models (VLMs) in multimodal fact-checking. While VLMs offer intrinsic fusion of visual and textual information, their effectiveness for veracity prediction remains contested. Several works report that end-to-end VLM predictions are sensitive to prompt design, model scale, and input modality balance, and may fail to reliably ground claims in retrieved evidence (Wang et al., 2025; Hetzner et al., 2025). Cekinel et al. propose a probing-classifier-based approach that extracts embeddings from VLMs and similarly to Hetzner et al. (2025) compares intrinsic multimodal embeddings against extrinsic fusion of separate text and image encoders (Cekinel et al., 2025). Their results suggest that explicit fusion of unimodal representations can outperform end-to-end VLM embeddings. Related findings in evidence-ranking and retrieval-augmented multimodal fact-checking further indicate that controlling how multimodal information is combined, rather than relying on implicit VLM fusion alone, is critical for robust verification performance (Tahmasebi et al., 2024). In contrast to representation-centric approaches, our work focuses on retrieval and reasoning control at the pipeline level. By decoupling question generation, modality-aware retrieval, and verdict prediction, we aim to leverage multimodal signals where they are most informative, particularly in query formulation and retrieval, without relying on heavy multimodal embedding fusion during evidence selection.

3 Methodology

3.1 Data

We conduct our experiments on the AVerImaTeC (Automated Verification of Image–Text Claim) dataset (Cao et al., 2025), which comprises 1,297 real-world fact-checked image–text claims split

[2]https://github.com/XplaiNLP/FEVER-9-AVerImaTeC-XxP

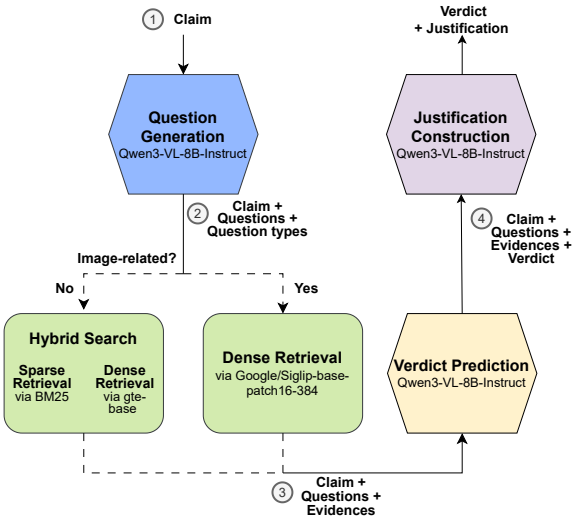

Figure 1: Architecture of the Proposed System.

into training, validation, and test sets. Each claim is associated with a set of verification questions and answers derived from web-based evidence, as well as textual justifications explaining the reasoning behind the final verdict. Claims are annotated with one of four labels: *Supported*, *Refuted*, *Not Enough Evidence*, or *Conflicting Evidence*. In addition to the annotations, the dataset provides a structured knowledge store containing scraped web text and images, which serves as the retrieval corpus for evidence collection.

3.2 Pipeline Design

Our system follows the overall architecture illustrated in Figure 1 and is inspired by the methodological framework of a competitive (5th place) submission to the AVeriTeC 2.0 Shared Task (Upravitelev et al., 2025). While retaining the core pipeline structure, we introduce targeted modifications to better accommodate the multimodal characteristics of image–text claim verification.

Given an input claim and its associated image, we first prompt a visual language model (VLM), namely *Qwen3-VL-8B-Instruct* [3] (Bai et al., 2025), to generate a set of three to five verification questions. These questions are designed to decompose the claim into verifiable sub-aspects and explicitly include at least one image-related question. Each question is labeled according to whether it pertains primarily to textual or visual evidence. The prompt used for question generation is provided in the appendix.

[3]https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

Evidence retrieval is then performed independently for each generated question, conditioned on its assigned type. For all questions, the retrieval query is formed by concatenating the original claim text with the generated question.

Text-related questions are handled using a hybrid retrieval pipeline adapted from Upravitelev et al. (2025). In this pipeline, the textual knowledge store is first segmented by concatenating consecutive sentences into fixed-length blocks of four sentences. Sparse retrieval is applied using *BM25* (Robertson and Zaragoza, 2009) to obtain a broad candidate set, from which the top 2,000 segments are retained. These segments and the query are subsequently encoded using the *gte-base* [4] embedding model (Li et al., 2023), and cosine similarity is computed to rank candidates. The top 10 most similar segments are selected as the final textual evidence.

Image-related questions are addressed using a separate multimodal retrieval pipeline based on *SigLIP* [5] (Zhai et al., 2023). The text-based, image-related knowledge store is first divided into token-aware chunks with a maximum length of 60 tokens. Each chunk is embedded using *SigLIP* [5] (Zhai et al., 2023), while the claim is represented by a joint embedding obtained through mean-averaging the textual query embedding and the embedding of the image referenced in the question. Cosine similarity is computed between the claim representation and all chunk embeddings. To obtain URL-level relevance scores, the top three chunk similarities per URL are averaged. The retrieval stage returns the top five URLs, along with the three most relevant text chunks for each URL.

The retrieved evidence from all questions is subsequently used to determine the final claim verdict. Conditioned on the claim, its associated image, and the collected evidence, the VLM (*Qwen3-VL-8B-Instruct* [3] (Bai et al., 2025)) is prompted to assign one of the four dataset labels.

Finally, the model is prompted to generate a concise natural-language justification consisting of two to four sentences that explains how the evidence supports the assigned verdict. The prompts used for verdict prediction and justification generation are included in the appendix.

The main idea follows the baseline approach of classifying each generated question by type, allowing us to apply a dedicated and more effective

[4]https://huggingface.co/thenlper/gte-base
[5]https://huggingface.co/google/siglip-base-patch16-384

retrieval strategy for each category. For questions concerning the textual component of the claim, evidence retrieval can leverage the efficient pipeline introduced by Upravitelev et al. (2025). Questions related to the image, however, require a separate retrieval pipeline. Because the text-based, image-related knowledge store is relatively small, we focused on dense retrieval methods for this component.

This design reflects our hypothesis that, under current retrieval capabilities and benchmark constraints, textual evidence offers more reliable grounding for verdict prediction than retrieved images, while visual information is most effective when used to guide question formulation and evidence discovery rather than direct veracity assessment. Importantly, this does not imply that images are uninformative in principle, but rather that current pipelines and evaluation protocols primarily reward text-grounded reasoning.

For multimodal retrieval, we selected *SigLIP*[5], which is pretrained on WebLI, a corpus containing real-world images and their corresponding captions (Zhai et al., 2023). We expect SigLIP to perform better than alternative models such as CLIP (Zhang et al., 2024) in this context, as it has been shown to outperform them on similar tasks, such as a zero-shot image-to-text retrieval on the MSCOCO dataset (Zhai et al., 2023; Chen et al., 2015).

4 Evaluation

Table 1 shows our results on the official FEVER-9 shared task leaderboards with regard to the validation[6] and test[7] set. The reported metrics are defined by Cao et al. and are presented below.

Question Score quantifies how well the system-generated questions align with the reference QA pairs used in the FEVER-9 pipeline. Because FEVER-9 evaluates question quality indirectly through the ability of the questions to retrieve appropriate evidence, this score reflects whether the generated questions adequately capture the semantic aspects required for verification.

Evidence Score measures evidence recall using the separated reference-based evaluation procedure. Each retrieved item is compared against canonical, LLM-generated evidence statements derived from the ground-truth QA pairs; a match is counted only when both the textual content and associated image tokens align with the reference. This provides an estimate of how effectively the system retrieves the information needed for accurate verification.

Verdict Accuracy is computed conditionally on evidence quality. Following the definition (Cao et al., 2025), a predicted verdict is considered correct only when the associated evidence score exceeds the predefined threshold $\lambda = 0.3$; otherwise, the prediction is counted as incorrect irrespective of its label. This formulation enforces the requirement that veracity assessments must be supported by sufficiently relevant evidence.

Justification Score evaluates the coherence and adequacy of the generated natural-language justifications using ROUGE-1 with respect to human-authored references. As with verdict accuracy, justification quality is also conditioned on evidence: any claim for which the evidence score falls below $\lambda = 0.3$ automatically receives a justification score of zero. This ensures that explanatory outputs are not only well-formed but also grounded in retrieved evidence.

Taken together, the results underscore that evidence retrieval is the primary determinant of system performance within the FEVER-9 framework. Our approach's substantial gains in evidence recall, despite its lower question similarity scores and its reliance solely on text-based retrieval, demonstrate that effective verification is driven less by producing questions that closely match reference annotations or by employing full multimodal fusion, and more by generating questions that reliably elicit semantically relevant information. This improved retrieval capability directly enables more accurate and well-supported verdicts and justifications under the evidence-conditioned evaluation protocol.

4.1 Ablation Study

A comparison with an earlier version of the system, summarized in Table 2, shows that the observed performance gains arise from changes across multiple components of the pipeline.

First, the initial submission relied exclusively on BM25 for evidence retrieval, whereas the current system employs an enhanced retrieval module that yields substantially higher evidence recall on the validation set (from 0.1188 to 0.2227). This improvement directly contributes to higher conditional verdict accuracy and justification scores.

Second, justification generation differed markedly between the two versions. The earlier system produced justifications using a minimal

[6]https://huggingface.co/spaces/FEVER-IT/AVerImaTeC
[7]https://huggingface.co/spaces/FEVER-IT/AVerImaTeCTest

Name	Split	Question Score	Evidence Score	Verdict Accuracy	Justification Score
FEVER-9 Baseline	dev	0.4882	0.1335	0.0658	0.0576
	test	0.5545	0.1707	0.1136	0.1322
XxP	dev	0.3729	0.2227	0.2303	0.175
	test	0.3902	0.2703	0.2557	0.198

Table 1: FEVER-9 Baseline and XxP (our) results on the AVerImaTeC.

Name	Split	Question Score	Evidence Score	Verdict Accuracy	Justification Score
XxP initial	dev	0.4720	0.1188	0.0987	0.011
XxP final	dev	0.3729	0.2227	0.2303	0.175

Table 2: XxP (initial) and XxP (final) results on the AVerImaTeC dataset.The initial version utilized BM25-only retrieval, template-based justifications and unstructured question generation. The final version implements the enhanced retrieval module (BM25 + SigLIP), instruction-based justification generation and modality-aware JSON question formatting. Both implementations employ an equivalent logical approach for verdict prediction.

template that simply restated the verdict and incorporated fragments of retrieved evidence. The final system replaces this approach with a structured, instruction-based prompt requiring concise, evidence-grounded reasoning, which results in justifications that align more closely with human-written references, reflected in the improvement from 0.0110 to 0.175.

Third, the question generation module in the initial system produced unstructured questions without explicitly indicating whether they referred to textual or visual information. In contrast, the updated version generates JSON-formatted questions that encode the question type (text-related vs. image-related), enabling a retrieval process that is more tightly aligned with the informational requirements of each claim. Although this change results in a decrease in Question Score, it leads to substantially higher evidence recall and, consequently, improvements in downstream verdict accuracy and justification quality. This result suggests that the Question Score metric may favor surface-level similarity to reference questions rather than the semantic usefulness of generated questions for evidence retrieval. In the FEVER-9 evaluation setting, where verdict and justification metrics are conditioned on evidence quality, improvements in retrieval effectiveness outweigh decreases in question similarity, highlighting a potential mismatch between question-level evaluation and end-to-end verification performance.

Overall, the ablation results indicate that improvements in retrieval effectiveness are the primary contributor to downstream performance gains, followed by more principled justification prompting and structured question generation. Taken together, these findings highlight that effective claim verification depends primarily on retrieving and reasoning over relevant evidence, rather than on surface-level question similarity or simplistic explanatory templates.

5 Discussion and Future Work

Future work spans several directions across question generation, verdict prediction, justification quality, and retrieval. For question generation, a promising avenue is fine-tuning models for question-type classification using the full training dataset, including explicitly learning to produce both text-related and image-related questions. Additional improvements may be achieved by experimenting with alternative prompting strategies, such as few-shot prompting, as well as by evaluating more powerful multimodal language models and systematically comparing different models for each subtask.

In the current system, final verdict prediction and justification generation rely exclusively on textual evidence. This design choice is hypothesis-driven: we assume that, under current retrieval capabilities and evaluation protocols, text-based evidence provides more stable grounding for veracity assessment. While this choice yields strong performance in our experiments, it also highlights an interesting direction for future comparison. In particular, incorporating retrieved images directly into the verdict prediction stage would enable controlled comparisons between text-grounded and image-grounded reasoning within an otherwise identical pipeline. Such comparisons could provide valuable insight into when visual evidence contributes decisively to claim verification, for example in cases involving

fine-grained visual details or image-based disambiguation.

Our proposed architecture outperforms the baseline in verdict accuracy, evidence quality, and justification score, indicating that retrieval effectiveness, particularly the choice and configuration of retrieval tools, plays a critical role in downstream verification performance. This suggests that further gains may be obtained by refining retrieval strategies and better aligning them with the informational requirements of different verification questions. In particular, the potential benefit of incorporating additional tools, such as web image search or visual question answering, depends critically on maintaining a well-balanced upstream tool-selection mechanism. Introducing new tools may introduce biases in the selection process, potentially degrading the primary objective of verdict accuracy.

Nonetheless, there are use cases in which enriched retrieval strategies could be highly beneficial, for example for claims requiring visual disambiguation or access to complementary sources of evidence. Future work should therefore focus on improving tool-selection methods to more effectively leverage heterogeneous retrieval capabilities. Moreover, it appears essential to optimize retrieval hyperparameters, such as the number of segments retrieved via BM25 and the number of chunks used for URL-level scoring, for specific use cases. Incorporating adaptive retrieval methods into the sparse retrieval stage may further improve the efficiency and performance of the pipeline (e.g., Rathee et al. (2025)). Overall, while the proposed architecture demonstrates strong performance under the current experimental setup, further refinement and systematic evaluation are required before deployment outside controlled laboratory settings.

6 Conclusion

In this work, we presented an efficiency-aware, selective multimodal fact-checking pipeline that treats multimodality as a controllable design choice rather than a uniformly applied system property. By decomposing claims into text-related and image-related verification questions and applying modality-specific retrieval strategies, our approach leverages visual information where it is most informative, primarily in guiding question formulation and evidence discovery, while relying on textual evidence for robust verdict prediction and justification generation.

Our evaluation on the AVerImaTeC benchmark demonstrates that this selective strategy achieves competitive performance, outperforming the FEVER-9 baseline in evidence recall, conditional verdict accuracy and justification quality, and ranking fourth overall in the shared task.

Limitations

Our work has several limitations related to computation, question generation, prediction, and retrieval. First, the overall system is constrained by computational and time limitations, which restricted the size of the models we could employ and limited the number of experimental configurations we were able to explore. In the question-generation component, imposing explicit structural constraints to classify questions as text-related or image-related reduced surface fluency and naturalness, reflecting a trade-off between expressive question formulation and structured control. While these constraints negatively affect question-level similarity metrics, they enable more precise alignment between generated questions and modality-aware retrieval strategies, which improves evidence recall and downstream verification performance. More broadly, our approach relies exclusively on zero-shot prompting, uses a single relatively small model across all stages, and is trained and evaluated using only the validation set rather than the full training data. These choices limit the expressiveness of the system and may affect its ability to generalize beyond the current experimental setup.

A further limitation concerns the scope at which visual information is incorporated into the pipeline. While the system deliberately uses images during question generation and modality-aware retrieval, final verdict prediction and justification generation are based solely on textual evidence. This is a hypothesis-driven design choice aligned with current retrieval capabilities and evaluation protocols, rather than a technical constraint. However, this choice limits our ability to directly assess the contribution of retrieved images to downstream veracity assessment. In particular, we do not evaluate potential scenarios in which visual evidence, such as alternative viewpoints, fine-grained visual details, or image manipulation, might provide additional verification signals beyond what is captured in text.

Addressing this limitation would require systematic comparisons between text-grounded and image-grounded verdict prediction within a shared

pipeline, as well as retrieval strategies that jointly consider images and their associated textual context. Such extensions are left to future work.

Acknowledgments

This research was carried out as part of the *VeraXtract* (reference: 16IS24066) and *news-polygraph* (reference: 03RU2U151C) projects, both supported by funding from the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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*.

Recep Firat Cekinel, Pinar Karagoz, and Çagrı Cöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of COLING*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *Preprint*, arXiv:1504.00325.

Michael Hameleers, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301.

Thomas Hetzner, Veronika Solopova, Vera Schmitt, and Dorothea Kolossa. 2025. Integrating video, text, and images for multimodal disinformation detection. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation (MAD'25)*. ACM.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025. Quam: Adaptive retrieval through lt;ugt;qult;/ugt;ery lt;ugt;alt;/ugt;ffinity lt;ugt;mlt;/ugt;odelling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 954–962. ACM.

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

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. 2024. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2189–2199, New York, NY, USA. Association for Computing Machinery.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*.

Max Upravitelev, Premtim Sahitaj, Arthur Hilbert, Veronika Solopova, Jing Yang, Nils Feldhus, Tatiana Anikina, Simon Ostermann, and Vera Schmitt. 2025. Exploring semantic filtering heuristics for efficient claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 229–237, Vienna, Austria. Association for Computational Linguistics.

Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. 2025. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. In *Proceedings of the Building Trust Workshop at ICLR*. ArXiv:2406.11288.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer.

A Appendix

A.1 Prompt for Question Generation

```
You are a professional fact-checker. Your task is to analyze a claim and the associated {Number of Images} images (if any), and generate a list of verification questions in a STRICT JSON format.

Each question must:
- Focus on a single, clear factual statement (atomic or meta-level).
- Be legitimate, clearly phrased, and directly related to the claim.
- Be verifiable using retrieved evidence (text or images).
- Be labeled as either text-related or image-related.

For EACH question, you MUST decide:
- "type": "text" - it is about ONLY the textual claim.
- "type": "image" - it is about AT LEAST one of the provided images.

Images are numbered in the order given: 1, 2, 3, ...

OUTPUT FORMAT (VERY IMPORTANT):
You MUST output ONLY a valid JSON array. No extra text, no comments, no explanation.

Example format (structure only, not content):
[
"id": 1,
"question": "Question text here",
"type": "text",
"images": []
,

"id": 2,
"question": "Question text here",
"type": "image",
"images": [1, 2]
]

Rules:
- Generate between 3 and 5 questions.
- You must NOT generate repeated questions.
- If there is at least one image, at least one question MUST be of "type": "image".
- The questions can be about different aspects of the claim.
- "id" must be 1, 2, 3, ... in order.
- If "type" is "text", "images" MUST be [].
- If "type" is "image", "images" MUST be a non-empty list of 1-based image indices.

Claim:
{Claim Text}

Now output ONLY the JSON array described above. Do NOT wrap it in backticks and do NOT add any text before or after it.
```

A.2 Prompt for Producing a Final Verdict

```
You are a professional fact checker.
Given a claim, evidence and images, classify the claim's veracity into EXACTLY ONE of:

- Supported
- Refuted
- Conflicting
- Not Enough Evidence

Definitions:
- Supported: The evidence clearly supports the claim.
- Refuted: The evidence clearly contradicts the claim.
- Conflicting: The evidence both supports AND refutes parts of
```

the claim.
- Not Enough Evidence: Evidence is insufficient, unrelated, or too weak to decide.

Answer with ONLY one of these four labels, nothing else.

Claim:
{Claim Text}

Evidence:
{Retrieved Evidences}

A.3 Prompt for Justification Generation

You are a professional fact checker.

Your task:
Given a claim, a FINAL veracity label, and the retrieved evidence (and optionally images), write a brief justification that explains why this label is appropriate.

Requirements:
- The justification MUST be consistent with the given label: "verdict".
- Base the explanation ONLY on the provided evidence (and images), do not invent facts.
- Refer explicitly to the key pieces of evidence.
- Be concise: 2–4 sentences.
- Do NOT restate the label; focus on the reasoning.

Claim:
{Claim Text}

Final label:
{Produced Verdict}

Evidence:
{Retrieved Evidences}