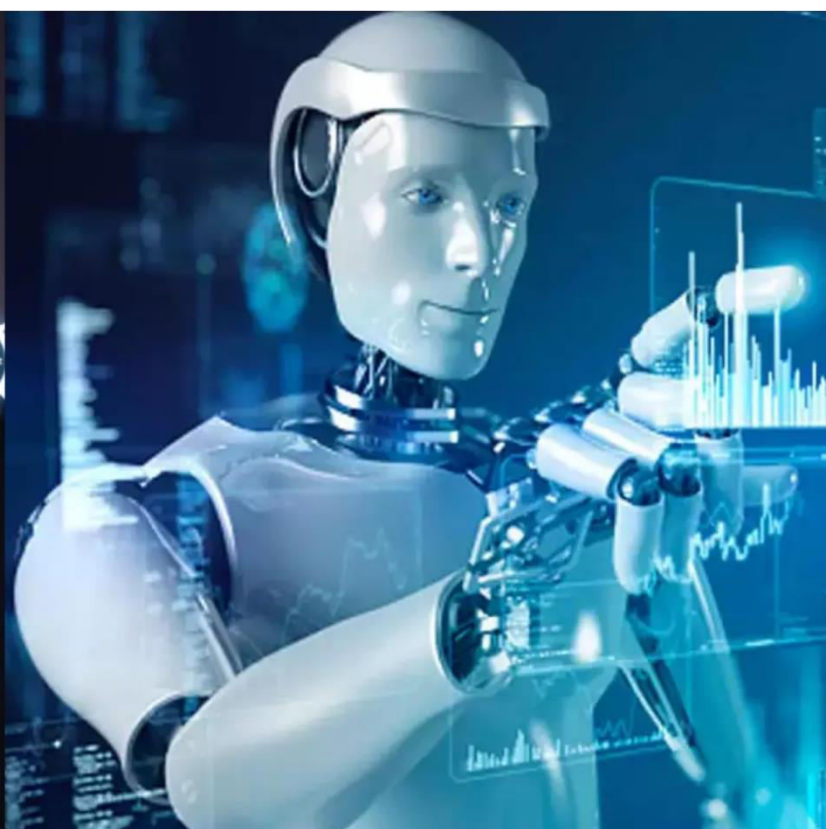




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 10, October 2026

An Explainable Multimodal Large Language Model Based Out of Context Image Caption Misinformation Detection

Raksha U. Agarkar, Om U. Dhokane, Kamlesh N. Dudhe, Govind R. Gulhane, Y. A. Dhumale

Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology & Research, Badnera,
Amravati, Maharashtra, India

ABSTRACT: Out-of-Context (OOC) disinformation couples a verified photo with a caption that misrepresents the event, location, persons or period to which it refers, such that pixel-level forensics cannot uncover it and image-text contrast is an intermediate signal. This paper introduces OOC-Verify, a verification workflow comprising a CLIP cosine similarity score combined with a logistic-regression alignment classifier, reasoning on claims by 4-bit-quantised Qwen2.5-VL-3B-Instruct, keyword-driven evidence search in Google News RSS feeds, and a Qwen-prioritised decision step producing a Prediction (Genuine, Misleading or Inconclusive), a heuristic confidence score, a Reason, Evidence articles, and an Explanation. On the entire NewsCLIPpings test set (N=7,264 pairs), the standalone CLIP-based classifier—being the sole pipeline component to be evaluated on such a scale—achieves 65.28% accuracy, ROC-AUC 0.7197, and F1 0.6392, with a 30.95% false-positive and 38.49% false-negative rate. An exploratory sample of 100 pairs yields 46.00% accuracy, inconsistent with the overall test performance and treated as an unresolved smoke test, not as evidence of generalisation. A controlled experiment involving 10 cases based on only two images gives accuracies of 60% (alignment-only), 100% (MLLM-only) and 100% (proposed pipeline); the proposed pipeline does not improve on MLLM-only, and the study is too small to support generalisation. Abstention and explanation quality were not assessed.

KEYWORDS: misinformation in out-of-context use, cheapfakes, multimodal large language models, CLIP, explainable AI, evidence discovery, decision fusion.

I. INTRODUCTION

Many instances of visual misinformation are not actually manipulated: a true photograph is described in the caption with a claim regarding what the image represents that is untrue, either regarding location, time, or other factors [1], [2]. Forensic techniques like error-level analysis or artifact detection cannot be used since the pixels have not been changed [2], [3], and pure text-based fact checkers overlook whether the image supports the claim [1].

A natural baseline is used to threshold CLIP [5] image-text cosine similarity. But misleading captions often share generic vocabulary with the image and obtain equal values; hence, scalar similarity differentiates the categories only partially (V). For example, a picture of the residential road flooded during an autumn storm of 2019 and the caption “Severe monsoon flooding paralyzes Mumbai in July 2025” share flooded asphalt, submerged cars and rain and therefore obtain overlapping embeddings due to the factual inconsistency. The previous literature also has the following weaknesses: many detectors give only binary decisions without an explanation [4], [6]; multimodal LLMs may just repeat the query rather than analyzing the pixel information [2], [7], [8]; separation of prediction, reasoning, evidence, and explanations is rare [1], [9]; and abstinence when neither an image nor evidence confirms the claim is infrequent [10].

OOC-Verify, our student research prototype that tackles these shortcomings in a transparent and rule-traceable manner, is presented. Our contributions include:

- 1) A complete evaluation test ($N = 7,264$) on the one-featured CLIP alignment classifier, including its error rates, along with the clear distinction of the findings from preliminary tests.
- 2) A seven-stage pipeline comprising the alignment classifier, 4-bit NF4-quantised Qwen2.5-VL-3B-Instruct reasoning (run on a laptop GPU with 6 GB memory), Google News RSS retrieval and a Qwen-prioritised decision step, with a remote OpenRouter fallback that was not used in the reported experiments.
- 3) A deterministic, Qwen-prioritised decision step: when Qwen returns Genuine or Misleading, its verdict fixes the binary prediction, while alignment and retrieved evidence shape only the heuristic confidence score and the Reason; an Inconclusive state is available when Qwen abstains or its output cannot be used. Abstention is not quantitatively evaluated here.
- 4) Controlled ablation study of ten cases; limitations included: fusion does not improve accuracy over MLLM reasoning alone.

II. RELATED WORK

Benchmarks and baselines. NewsCLIPpings [4] builds OOC pairs from VisualNews [11] by matching images with captions of other articles using CLIP, SBERT person and ResNet scene retrieval; its baselines are binary classifiers without explanations or retrieval. COSMOS [6] detects OOC use via self-supervised caption–image grounding, and Abdelnabi et al. [7] verify image–caption pairs against online evidence. Surveys [1], [9], [12] identify explainability and evidence grounding as open bottlenecks.

MLLM- and evidence-based methods. Multimodal LLMs build on visual instruction tuning [8], and Qwen2.5-VL [16] is the model used here. SNIFFER [2] instruction-tunes a multimodal LLM with external knowledge for explainable OOC detection on NewsCLIPpings and Twitter-COMMs. EXCLAIM [13] uses an agentic system with hierarchical retrieval, and CMIE [14] combines MLLM insights with external evidence to counter MLLM hallucination. COVE [10] and Tonglet et al. [3] predict the original context of an image to judge veracity. Duwal et al. [15] aggregate captions, image entities, and retrieved documents in an attention-based graph neural network, which outputs labels without natural-language explanations. Table I summarises the closest prior work.

TABLE I CLOSEST PRIOR WORK AND ITS RELATION TO OOC-VERIFY

Work	Approach	Relation to this work
NewsCLIPpings [4]	OOO pairs generated with CLIP, SBERT and ResNet retrieval; balanced partition of 85,360 pairs	Benchmark; source of our training and evaluation splits
VisualNews [11]	1.3 M news image–caption pairs from four agencies; no misinformation labels	Source corpus of the authentic pairs
COSMOS [6]	Self-supervised caption–image grounding	Label-only output; no rationale
SNIFFER [2]	Two-stage instruction-tuned MLLM with external knowledge	Inspires our prompt structure and entity decomposition
EXCLAIM [13]	Agentic cross-modal system with hierarchical retrieval	Motivates the staged, evidence-structured pipeline
CMIE [14]	MLLM insights combined with external evidence	Supports hybrid visual + retrieval fusion
COVE [10]	Predicts the original context to judge veracity	Motivates separating contextual from visual claims
Duwal et al. [15]	Attention-based GNN over captions, entities and retrieved evidence	Evidence-grounded; labels only, no explanations

Positioning. OOC-Verify is a lightweight, rule-based composition of a CLIP classifier [5], an off-the-shelf MLLM [16] and keyword retrieval over Google News RSS [17] that exposes each intermediate signal and returns four deliverables: a prediction, a reason, linked evidence and an explanation. We make no empirical comparison with the systems above and do not claim superiority over them.

III. METHODOLOGY

A. Problem formulation and taxonomy

An input is a pair of photo-caption $P = (I, C)$, where I is a genuine photo. The output is a decision $Y^* \in \{\text{Genuine, Misleading, Inconclusive}\}$ along with some additional information: heuristic confidence c^* , a principal reason r^* , a set of articles found E and a textual justification e^* . We distinguish four types of phenomena, to ensure that heuristics would not be confused: (1) visual inconsistency, which is an apparent contradiction between the visual content and the scaption (e.g. a green vs. red shirt), resolvable without any need for the external articles; (2) contextual misinformation, assertions about occurrences, locations, dates, or people that cannot be checked based on the pixel information alone; (3) external evidence relation, articles found supporting or refuting a contextual assertion; and (4) epistemic uncertainty, where neither pixels nor external articles can resolve the assertion and a forced binary decision might result in unjustified accusations, so the system can return Inconclusive. Table II lists the notation used below.

TABLE II MAIN NOTATION

Symbol	Meaning	Range
I, C	Input image and caption	–
u_I, u_C	L2-normalised CLIP image/text embeddings	$\mathbb{R}^{512}$, unit norm
S_{align}	CLIP cosine similarity	$[-1, 1]$
$y_{\text{align}}, \text{conf}_{\text{align}}$	Alignment prediction; posterior of predicted class	$\{\text{Gen.}, \text{Mis.}\}; [0.5, 1]$
y_{mlm}	Parsed Qwen2.5-VL prediction	$\{\text{Gen.}, \text{Mis.}, \text{Incon.}\}$
$S_{\text{supp}}, S_{\text{contra}}$	Evidence support/contradiction indices	$[0, 1]$
Y^*, c^*, r^*, e^*	Verdict, heuristic confidence index, reason, c^* heuristic, uncalibrated explanation	

Gen. = Genuine; Mis. = Misleading; Incon. = Inconclusive.

B. Pipeline overview

Figure 1 presents the seven stages that operate in sequence. React dashboard and FastAPI backend provide access to the pipeline, with the API receiving input files in the form of JPEG/PNG/WebP images up to 10MB in size. In the case of local execution, all stages are run locally; in the case of remote execution, the MLLM query is sent to the OpenRouter models, while the confidence index is output as a placeholder 0.0 (uncalibrated). Remote execution was not evaluated in any of the evaluations carried out. Figure 2 shows the dashboard during a run, with the seven stages reported as they complete.

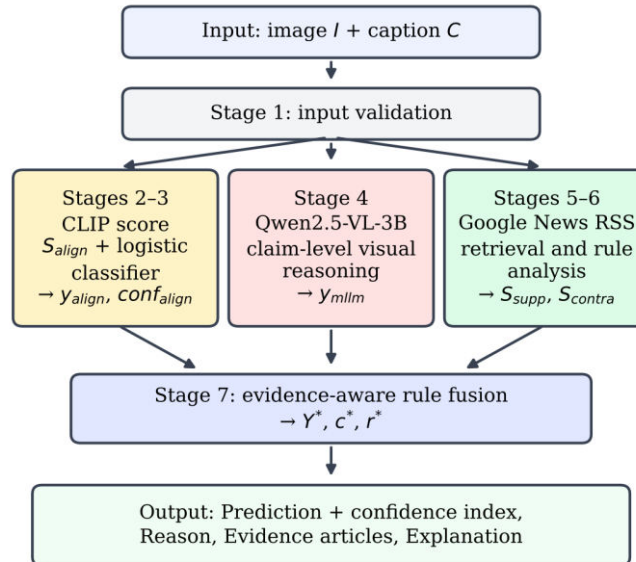
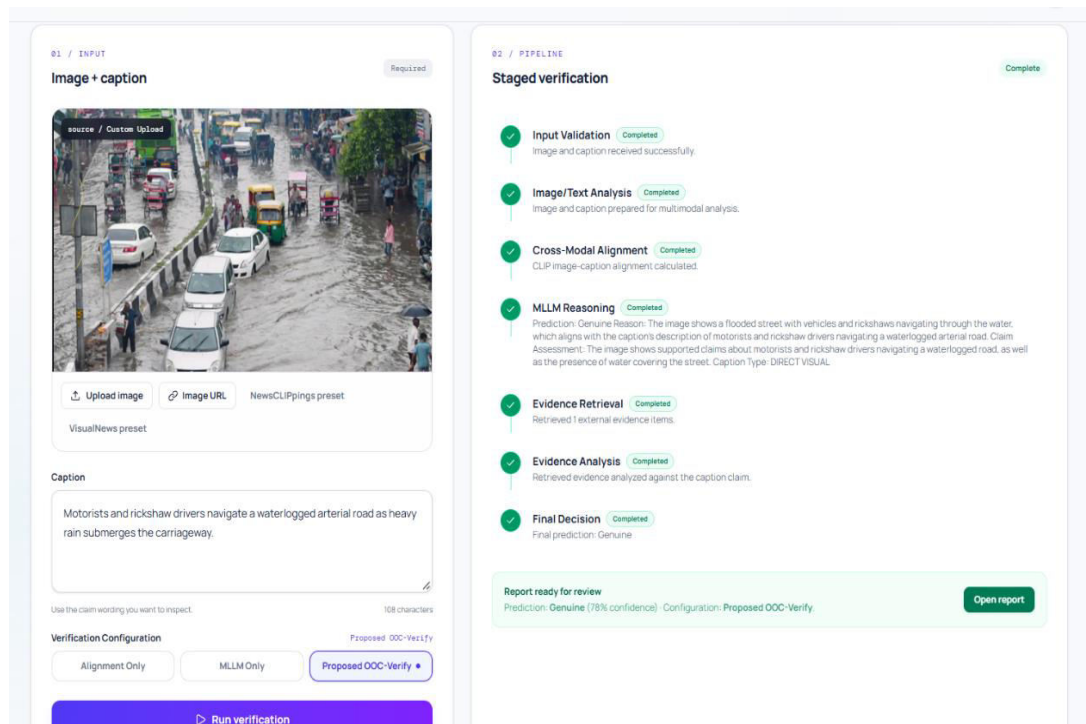


Fig.1. Pipeline for OOC-Verify (Stages 1-7, performed serially). Stage 1 verifies the caption and image. Stage 7 is Qwen-prioritised (Section III-F): alignment and evidence scores shape the confidence index and Reason but do not flip a Genuine or Misleading Qwen verdict.



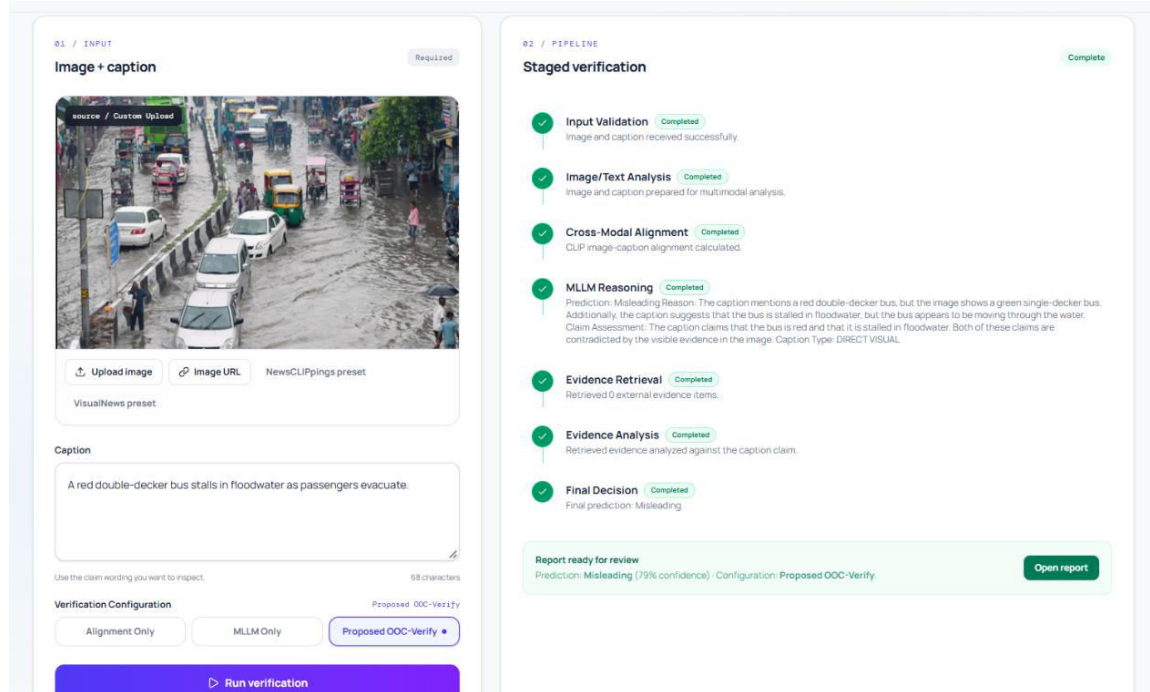


Fig.2. OOC-Verify dashboard (Verify view) for custom-upload image–caption pairs: Genuine case (top) and Misleading case (bottom).

C. Cross-modal alignment and categorization (Stages 2-3)

The image is transformed to RGB, resized to a $3 \times 224 \times 224$ tensor, and the caption is limited to 77 tokens. The output vector dimension of CLIP ViT-B/32 (openai/clip-vit-base-patch32) [5] is 512, and this vector is L2-normalized to u_I and u_C :

$$S_{\text{align}} = u_I \cdot u_C, \quad u = v / \|v\|_2 \quad (1)$$

A logistic regression analysis of scikit-learn on this single variable, using the NewsCLIPpings training data embeddings (Class 1 = Misleading), produces

$$P(\text{Misleading} | S_{\text{align}}) = 1 / (1 + \exp(-(b + w S_{\text{align}}))) \quad (2)$$

where the parameters have been learned with $w = -19.384933$ and $b = 5.647668$, meaning that $t = -b/w = 0.2913$; a sample is labeled as Genuine if $S_{\text{align}} \geq t$ and Misleading otherwise. The confidence of alignment, defined as $\text{confalign} = \max(P, 1-P)$, is the posterior probability of the labelled class.

D. MLLM reasoning (Stage 4)

Qwen2.5-VL-3B-Instruct [16] (Qwen/Qwen2.5-VL-3B-Instruct) is loaded in the active configuration with 4-bit NF4 quantisation through BitsAndBytes, with double quantisation enabled and FP16 compute, with device_map="auto" and memory caps of 4.0 GiB (GPU) and 4.5 GiB (CPU), so that it runs on a 6 GB RTX 3050 Laptop GPU. No disk offload is configured in this mode, and the caps refer to loader limits and not measured memory usage. An FP16 fallback configuration exists but is inactive and was not used in any reported experiment. Before inference, the image is downsampled with its aspect ratio preserved, so that the longer edge is at most 1,024 px and the area at most 1,003,520 px; images are never upsampled. This reduces the number of visual input tokens and the inference workload; it can discard fine detail, and we do not claim that it leaves image quality unaffected. Generation uses max_new_tokens=150, and no other decoding parameter (e.g., temperature or top-p) is set explicitly. The structured prompt instructs the model to (i) describe only the visible aspects, (ii) classify caption claims as either DIRECT_VISUAL or CONTEXTUAL, (iii) categorize each claim as SUPPORTED, CONTRADICTED or UNVERIFIED, noting that non-visible visual evidence

does not imply a false claim, and (iv) provide an answer in five lines (Prediction, Reason, Visual Evidence, Claim Assessment, Caption Type) that is subsequently classified into $y_{mlm} \in \{\text{Genuine, Misleading, Inconclusive}\}$. This five-line format, including Visual Evidence, is part of the Qwen-stage output carried within the pipeline response; it is distinct from the user-facing deliverables in Section III-G.

E. Data searching and interpretation (Stage 5-6)

The boilerplate is stripped away from the caption, and the query is restricted to the first 16 key terms from the caption. Purely visual captions do not retrieve. The query is submitted to Google News RSS [17]; at most $k = 5$ articles (title, date, source, snippet, URL) are stored, using a fallback proper noun query if there are no results. Each article a is scored based on keyword similarity, and if the caption contains any numbers, number similarity:

$$\text{rel}(a) = 0.75 \cdot |W_C \cap W_a|/|W_C| + 0.25 \cdot |N_C \cap N_a|/|N_C| \quad (3)$$

otherwise the keyword term itself. The following order is followed for testing relations: contradictions if $\text{rel} \geq 0.60$ with explicit contradictory terms; supports if $\text{rel} \geq 0.60$ (or 0.70 with exact event matches) with support patterns; irrelevant if $\text{rel} < 0.20$; otherwise insufficient. Mean relevance over matching terms is used for both support and contradiction indexes,

$$S_{\text{supp}} = (1/|E_{\text{supp}}|) \sum_{a \in E_{\text{supp}}} \text{rel}(a), \quad S_{\text{contra}} \text{ analogous} \quad (4)$$

with $S = 0$ for an empty set. Two contradiction thresholds therefore act at different levels: 0.60 applies to the relevance of an individual article when it is assigned the contradicts relation, whereas 0.70 applies to the aggregate index S_{contra} (a mean over articles) when fusion decides whether contradicting evidence is strong (Section III-F).

F. Qwen-prioritised decision fusion (Stage 7)

Fusion is Qwen-prioritised (Table III). When Qwen returns Genuine or Misleading, that verdict determines the final binary prediction; alignment (y_{align} , $\text{conf}_{\text{align}}$) and retrieved evidence (S_{supp} , S_{contra}) do not flip it. They affect only the heuristic confidence score and the selection of the Reason. When Qwen returns Inconclusive, or its output is unavailable or cannot be parsed, the system can return Inconclusive. Contradicting evidence counts as strong when the aggregate S_{contra} reaches 0.70, a threshold distinct from the 0.60 per-article threshold of Section III-E. The exact confidence formulas are not reproduced here because they have not been checked against the implementation for this draft; the confidence index is a heuristic, uncalibrated score and not a calibrated probability.

TABLE III QWEN-PRIORITISED DECISION LOGIC (STAGE 7)

Qwen output (y_{mlm})	Final prediction	Role of alignment and retrieved evidence
Genuine	Genuine	Affect the heuristic confidence score and Reason selection only; they do not flip the verdict
Misleading	Misleading	Affect the heuristic confidence score and Reason selection only; they do not flip the verdict
Inconclusive	Inconclusive	Qwen abstention is returned as Inconclusive
Unavailable or unparseable	Can be Inconclusive	No Qwen verdict is available to prioritise

The confidence index is a heuristic computed from alignment and evidence signals; it is uncalibrated and is not a probability of real-world veracity.

G. Outputs and explainability

The system returns the following deliverables (Table IV): the Prediction (with its heuristic confidence index), a Reason, the retrieved Evidence articles with relevance scores, relations and links, and a multi-sentence Explanation that traces which signals led to the verdict. Visual Evidence is not a separate top-level response field; it is a line of the Qwen-stage output carried within the pipeline-stage output. The Reason is selected during fusion from the Qwen verdict and the alignment and evidence signals, and the Explanation is generated text; their faithfulness and usefulness were not evaluated with users. The confidence index is a heuristic that has not been calibrated and must not be read as a

probability. Table IV summarises content and does not enumerate the exact JSON field names. Figure 3 shows how these deliverables appear in the report view.

TABLE IV DELIVERABLES RETURNED FOR EACH INPUT (CONTENT SUMMARY)

Deliverable	Content
Prediction	$Y^* \in \{\text{Genuine, Misleading, Inconclusive}\}$ with heuristic confidence index c^* (uncalibrated; placeholder 0.0 in remote mode)
Reason	Short text selected during fusion from the Qwen verdict and the alignment and evidence signals
Evidence	Up to five articles: title, source, date, link, relevance $rel(a)$ and relation tag (supports/contradicts/insufficient/irrelevant)
Explanation	Generated multi-sentence trace of the visual findings, evidence relations and alignment score behind the verdict

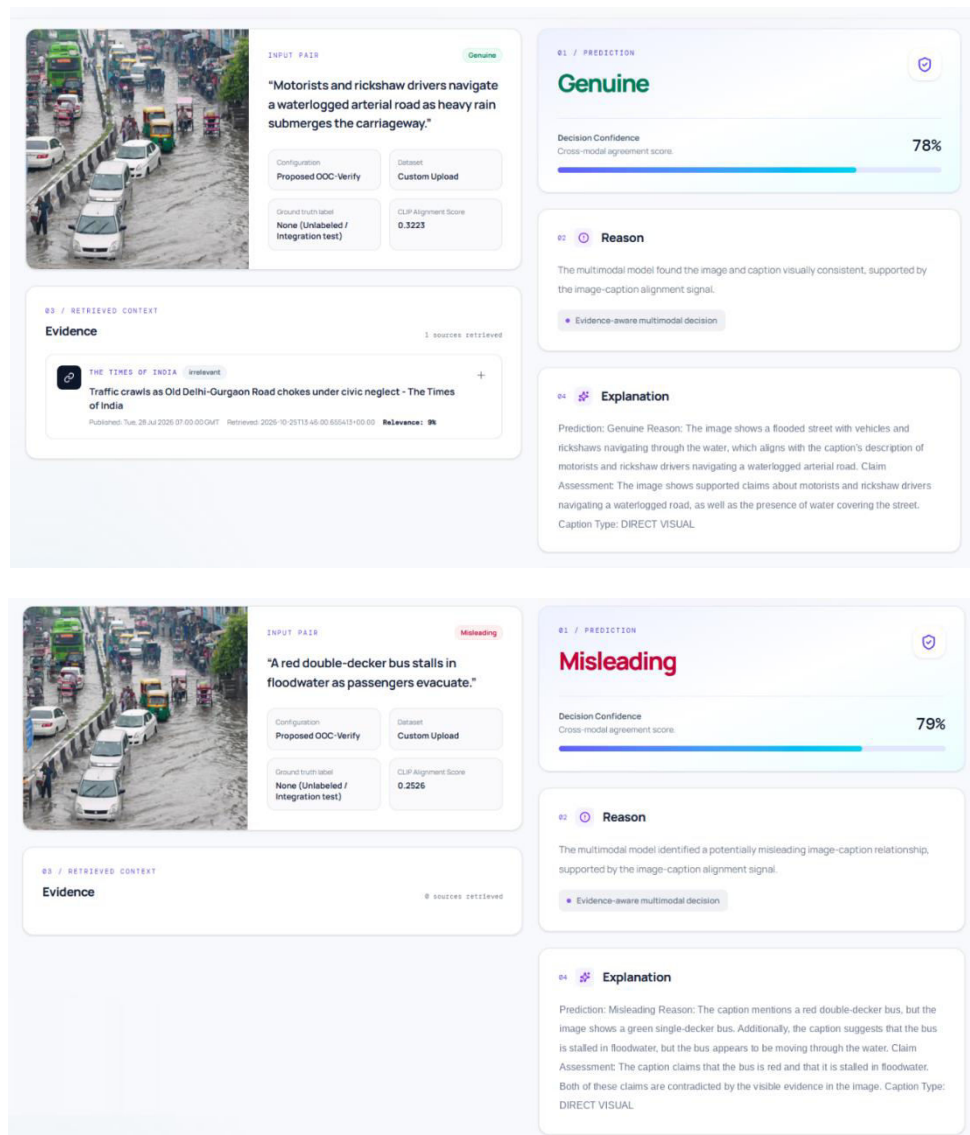


Fig.3. Report view showing the Prediction, Reason, Evidence and Explanation for a Genuine case (top) and a Misleading case (bottom). The confidence value is a heuristic index, not a calibrated probability.

IV. RESULTS AND DISCUSSION

Scope. Experiment 1 evaluates only the standalone CLIP alignment classifier (Stages 2–3). Only that component was run at scale; the end-to-end pipeline was run on ten cases.

Experimental protocol. The controlled set has ten cases (4 Genuine, 6 Misleading) built from two images, and Misleading is the positive class throughout. Three configurations were compared: alignment only (Stages 2–3), mllm_only (Stage 4) and proposed (the full seven-stage pipeline). Each configuration ran on all ten cases, giving 30 requests, and the run produced no Inconclusive predictions, so no mapping of Inconclusive outputs to the binary matrix was needed. The reported run used a local-only protocol on a Lenovo LOQ laptop (AMD Ryzen 7 7840HS, NVIDIA RTX 3050 Laptop GPU with 6 GB VRAM, 16 GB RAM): Qwen ran locally in the active 4-bit NF4 mode and OpenRouter was disabled. [Author confirmation needed: image provenance, licensing, the labelling protocol, and whether any rule or threshold was adjusted after the ten cases were seen.]

A. Full test split (N = 7,264)

TABLE V STANDALONE CLIP CLASSIFIER ON THE FULL TEST SPLIT

Metric	Value	Metric	Value
Accuracy	65.28%	Specificity	69.05%
ROC-AUC	0.7197	FPR	30.95%
Precision (Mis.)	66.53%	FNR	38.49%
Recall (Mis.)	61.51%	F1 (Mis.)	0.6392
TN / FP	2,508 / 1,124	FN / TP	1,398 / 2,234

Total errors: 2,522 of 7,264. The 95% Wilson interval for accuracy (4,742/7,264) is 64.18%–66.37%.

The classifier is clearly above chance (Table V), but one in three authentic pairs is flagged (1,124 false positives), and 38.49% of OOC pairs are missed (1,398 false negatives). The test split is exactly balanced and not representative of real prevalence, and NewsCLIPPings negatives selected by CLIP similarity are adversarial to a CLIP baseline; results were not broken down by negative-selection strategy. We hypothesise that OOC captions borrowed from similar events keep topical overlap with the image (false negatives) and that indirect or stylistic genuine captions lower similarity (false positives); no error analysis was run to test this.

B. Exploratory 100-pair subset

On the exploratory validation sample (Table VI; TN = 32, FP = 18, FN = 36, TP = 14), accuracy was 46.00% (Wilson 36.6%–55.7%), ROC-AUC 0.4928, macro-F1 0.4419, and mean scores were nearly identical (Genuine 0.3032, Misleading 0.3025). Under an exact one-sided binomial calculation that treats pairs as independent, an accuracy this low has probability about 6×10^{-5} if the true accuracy were 65.28%. We therefore do not read it as sampling variance; it is reported only as a pipeline smoke test and is evidence neither about CLIP nor of successful generalisation. Its cause (e.g., validation rather than test data, mismatched pairs in the subset file, a different preprocessing path or classifier version) is unresolved.

TABLE VI EXPLORATORY 100-PAIR SUBSET (PIPELINE SMOKE TEST)

Class (support 50)	Precision	Recall	F1
Genuine	0.4706	0.6400	0.5424
Misleading	0.4375	0.2800	0.3415
Macro average	0.4540	0.4600	0.4419

TN = 32, FP = 18, FN = 36, TP = 14; accuracy 46.00%; FPR 36.00%, FNR 72.00%.

C. Ten-case controlled ablation

TABLE VII ABLATION ON THE TEN CONTROLLED CASES (4 GENUINE, 6 MISLEADING; TWO IMAGES; MISLEADING = POSITIVE)

Configuration	Acc.	Prec.	Rec.	F1	TN	FP	FN	TP
Alignment-only	0.60	0.60	1.00	0.75	0	4	0	6
MLLM-only	1.00	1.00	1.00	1.00	4	0	0	6
Proposed	1.00	1.00	1.00	1.00	4	0	0	6

Wilson 95% intervals: 10/10 gives 72.2%–100%; 6/10 gives 31.3%–83.2%. Per-case outputs are not tabulated here.

Table VII reports the three configurations over 30 requests (10 cases $\times$ 3 configurations), with zero failed requests, zero missing predictions and zero Inconclusive predictions. Alignment-only predicted Misleading for all ten cases (TN = 0, FP = 4, FN = 0, TP = 6), so its 60% accuracy equals an always-Misleading baseline. MLLM-only and the proposed pipeline both reach 100% (TN = 4, FP = 0, FN = 0, TP = 6). The proposed pipeline therefore ties MLLM-only, and **it does not improve accuracy over MLLM-only on this small test**; no claim is made that fusion is superior. This tie is consistent with the design, because Qwen’s Genuine or Misleading verdict fixes the final prediction. Nine of ten cases are direct visual claims that the MLLM resolves from pixels; only two images are used, so cases are not independent, and the intervals are wide; the comparison is descriptive and does not support generalisation. The alignment-only result also reflects a domain shift: the classifier was trained on NewsCLippings and is evaluated here on short, manually constructed captions, so its 60% is not a measure of CLIP-based detection in general. On C-06, the only news caption, Qwen predicted Misleading directly; the out-of-context safeguard was not triggered, and we do not claim that the retrieved articles corroborated or contradicted the captioned event. What the full pipeline adds, namely linked evidence articles for human review, is not reflected in accuracy. No case expects Inconclusive, so abstention is untested, and no abstention rate or error reduction has been measured. Recorded Qwen generation times were 5.26–15.98 s per sample; these are generation latency only, not total API or pipeline latency, come from a single run, and do not constitute a reproducible latency benchmark or a throughput estimate.

D. Discussion

RQ1: scalar CLIP similarity gives moderate but insufficient discrimination. RQ2: on the nine direct-visual cases, the MLLM resolved every contradiction, which is consistent with, but too small to establish, a general advantage over alignment alone. RQ3: not established; the ten cases contain a single news case, which Qwen classified directly, and show no benefit of fusion. Because an automated “Misleading” label on authentic reporting can suppress legitimate content, the Inconclusive state is intended to reduce unsupported accusations, and all evidence articles are linked for human inspection; the abstention rate and its effect on errors have not been measured.

E. Limitations

- Only the CLIP baseline was evaluated on all 7,264 pairs; the end-to-end pipeline was run on only ten cases built from two images, so no claim is made about its accuracy on the benchmark or about generalisation.
- The 100-pair result is exploratory, conflicts with the full-test result, and its cause is unknown.
- Confidence indices are uncalibrated heuristics. Evidence relations come from keyword rules, retrieval is text-only and can return irrelevant articles, evidence does not flip a Genuine or Misleading Qwen verdict, and there is no reverse-image provenance search.
- The alignment classifier was trained on NewsCLippings and evaluated on short, manually constructed captions, a domain shift; the 60% alignment-only accuracy should be read accordingly.
- No user study was performed on the Reason or Explanation outputs.
- The test split is exactly 50:50, shares images across pairs, and no train/validation/test overlap check is reported; intervals should be recomputed by cluster or paired bootstrap.
- The implementation and evaluation files are not yet publicly available.
- The decision logic is hand-designed and Qwen-prioritised, so Qwen errors propagate to the final verdict. The confidence formulas, image provenance, licensing and the labelling protocol still need author confirmation, and Qwen generation latency was recorded in a single run only.

V. CONCLUSION

OOC-Verify combines a CLIP alignment classifier, 4-bit Qwen2.5-VL reasoning, news retrieval and a Qwen-prioritised decision step to return a prediction, a heuristic confidence index, a reason, linked evidence and an explanation. The only large-scale result is the CLIP baseline (65.28% accuracy, ROC-AUC 0.7197 on 7,264 test pairs), which shows moderate discrimination and substantial error rates. In a ten-case, two-image study, alignment-only reached 60% accuracy and both MLLM-only and the proposed pipeline reached 100%; the proposed pipeline matches, but does not improve on, MLLM-only reasoning, and the study is too small to support generalisation. Future work should (i) run the full pipeline on the complete test split, (ii) add cases that require abstention and use independent images, (iii) resolve the 100-pair discrepancy, (iv) replace keyword evidence analysis with an NLI cross-encoder, (v) add reverse-image provenance search, (vi) calibrate the confidence index, and (vii) evaluate explanations with users.

REFERENCES

- [1] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, "Multimodal automated fact-checking: A survey," in Findings of ACL: EMNLP 2023, 2023, p. 5430–5448.
- [2] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, "SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection," in Proc. IEEE/CVF CVPR, 2024, pp. 13052–13062.
- [3] J. Tonglet, M.-F. Moens, and I. Gurevych, "Image, tell me your story! Predicting the original meta-context of visual misinformation," in Proc. EMNLP, 2024, pp. 7845–7864.
- [4] G. Luo, T. Darrell, and A. Rohrbach, "NewsCLIPpings: Automatic generation of out-of-context multimodal media," in Proc. EMNLP, 2021, pp. 6801–6817.
- [5] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.
- [6] S. Aneja, C. Bregler, and M. Nießner, "COSMOS: Catching out-of-context image misuse using self-supervised learning," in Proc. AAAI, vol. 37, no. 12, 2023, pp. 14084–14092.
- [7] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in Proc. IEEE/CVF CVPR, 2022, pp. 14940–14949.
- [8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Adv. NeurIPS, vol. 36, 2023, pp. 34892–34916.
- [9] G. Cai, Z. Qiu, G. Bi, and Q. Liu, "A comprehensive survey of multimodal fake news detection: Datasets, methods, and challenges," Trans. Artif. Intell., vol. 2, no. 1, pp. 131–160, May 2026.
- [10] J. Tonglet, G. Thiem, and I. Gurevych, "COVE: Context and veracity prediction for out-of-context images," in Proc. NAACL-HLT, 2025, pp. 2029–2049.
- [11] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual News: Benchmark and challenges in news image captioning," in Proc. EMNLP, 2021, pp. 6761–6771.
- [12] T. Hu et al., "Beyond words and pictures: A survey of multimodal fake news detection," Expert Systems, vol. 43, no. 6, p. e70260, Jun. 2026.
- [13] Y. Wu, Z. Zhang, F. Wang, Y. Luo, H. Xiong, and N. Tang, "EXCLAIM: An explainable cross-modal agentic system for misinformation detection with hierarchical retrieval," arXiv preprint arXiv:2504.06269, 2025.
- [14] F. Li, J. Wu, C. He, and W. Zhou, "CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection," in Findings of ACL 2025, 2025, pp. 9342–9354.
- [15] S. Duwal, M. N. S. Shopnil, A. Tyagi, and A. M. Prama, "Evidence-grounded multimodal misinformation detection with attention-based GNNs," arXiv preprint arXiv:2505.18221, 2025.
- [16] S. Bai et al., "Qwen2.5-VL technical report," arXiv preprint arXiv:2502.13923, 2025.
- [17] Google LLC, "Google News RSS search documentation and feed specifications," 2024. [Online]. Available: <https://news.google.com/rss>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details