Does Multimodal Retrieval Reduce Hallucination in News Summarization? A Text-vs-Pixel RAG Study

Yuhao Cao
Department of Electrical & Computer Engineering
University of Toronto
davidyh.cao@mail.utoronto.ca

Kaiwen Yang
Department of Electrical & Computer Engineering
University of Toronto
kaiwen.yang@mail.utoronto.ca

Eric Wang
Department of Electrical & Computer Engineering
University of Toronto
hbo.wang@mail.utoronto.ca

Abstract

We study whether retrieving actual image evidence, not just text, reduces hallucination in LLM-based news summarization beyond a strong text-only retrieval-augmented generation (RAG) baseline. We build three systems that share one generator and prompt and differ only in retrieved evidence: B_1 (text passages), M (text plus fused image retrieval and captions), and M_{vision} (M plus the retrieved image pixels themselves), evaluated on a duplicate-audited, group-safe split of 1,023 BBC News articles with a claim-level LLM-judge faithfulness metric. Retrieval of any kind is the dominant effect, cutting hallucination from 85% (no retrieval) to roughly 10–14%. Adding images is directionally positive at every step of the ladder but statistically null at the two prespecified comparisons, and only borderline significant, uncorrected, for the full text-to-pixel jump ($p = 0.057$). Targeted diagnostics show pixels are not inert: they measurably help under visually salient evidence and measurably hurt under image-text conflict, but this effect does not survive averaging over an unfiltered test role at the current sample size.

Attestation of Teamwork

Member	Contribution
Yuhao	System architecture: designed and built the retrieval-generation pipeline, wrote and refine the report
Kaiwen	Expanded the experiment suite and prepared the data, wrote and refine the report
Eric	Refined the evaluation methodology, developed the theoretical background, wrote and refine the report

ECE 1508: Deep Generative Models (Summer 2026)

1 Introduction

Large language models summarize fluently but are prone to hallucinating facts absent from the source text, a failure mode retrieval-augmented generation (RAG) [1] is designed to mitigate by grounding generation in retrieved evidence. News articles are naturally multimodal: a headline and body are usually paired with a photograph that sometimes carries information the text does not state explicitly. This raises a specific, testable question: does retrieving and conditioning on actual image evidence reduce hallucination beyond what a strong text-only retriever already achieves? We answer this empirically. The generator, prompt, and decoding are held fixed and only the retrieved evidence varies, isolating the causal effect of retrieval modality. We evaluate on a real corpus (BBC News, January 2024) with a duplicate-audited, group-safe train/development/test split, scored by a claim-level LLM-judge faithfulness metric rather than a surface-overlap metric that cannot detect hallucination.

2 Preliminaries and Problem Formulation

Given a query q and a corpus $\mathcal{C}$ of article-image pairs, a RAG system retrieves evidence $E \subset \mathcal{C}$, $|E| = k$, and generates a summary $y = G(q, E)$ with a fixed generator G. We define **faithfulness** as the fraction of atomic claims in y supported by E (grounded in the retrieved evidence, not necessarily true in the world), decomposed and verified by an LLM judge blind to which system produced y; **hallucination** is $1 -$ faithfulness. The question decomposes into a system ladder: does retrieval at all reduce hallucination relative to none (B_0)? Does adding an image retrieval stream and captions (M) help beyond text alone (B_1)? Does giving the generator the actual retrieved pixels (M_{vision}) help beyond captions? Answering this required three pieces of machinery new to this project: *late score fusion* across two embedding spaces (Section 4.1), a *retrieval-confidence abstention gate*, and a *decompose-then-verify* claim-level judge with paired bootstrap and Wilcoxon significance testing in place of a single point estimate.

3 Design

Table 1 summarizes the six-stage pipeline. Only the retrieval stage differs between B_1, M, and M_{vision}; every other stage (generator, prompt, decoding, and judge) is shared, so a difference between systems can be attributed to the evidence each one receives rather than to an implementation difference.

Stage	Implementation
Data	1,023 BBC News article–image pairs; duplicate-audited, group-safe 707/166/150 pool/development/final-test split
Embedding	SBERT `all-MiniLM-L6-v2` (text, 384-d) [3] and CLIP `ViT-B/32` (image & query, 512-d) [2]; both frozen, pretrained
Indexing	Two FAISS `IndexFlatIP` indexes [4] on unit-normalized vectors (exact cosine search)
Retrieval	Late score fusion of the two streams (Section 4.1); the system's only independent variable
Generation	One `generate(prompt)` call to `gpt-4o-mini` [7], T=0, byte-identical across arms except the evidence block; retrieval-confidence abstention gate
Evaluation	Recall@k against the withheld gold article; decompose-then-verify LLM-judge faithfulness [5,6], paired bootstrap + Wilcoxon

Table 1: The six pipeline stages.

All retrieval and prompt choices (α, k, reranking) are selected on the development role only and frozen in `configs/final_research.json` before the final-test role is touched, so the reported numbers cannot be an artifact of tuning on the test set.

4 Methodology

4.1 Late Score Fusion

Given a query q and a candidate article set $\mathcal{C}$, the two retrieval streams are scored independently. An article is scored by its best-matching passage, so that a long article is not penalised for containing unrelated text:

$$s_{\text{text}}(a) = \max_{p \in a} \cos\big(f_{\text{SBERT}}(q),\, f_{\text{SBERT}}(p)\big) \tag{1}$$

For the image stream the same query q is embedded with CLIP's text encoder, which places it in the shared image–text space and lets a query q search an index of images:

$$s_{\text{img}}(a) = \cos\big(f_{\text{CLIP-T}}(q),\, f_{\text{CLIP-I}}(x_a)\big) \tag{2}$$

The two streams occupy different score ranges, so a weighted sum of the raw values would not be a weighted average of anything comparable. Each stream is therefore min–max normalized *within the candidate set of the current query*,

$$\hat{s}(a) = \frac{s(a) - \min_{a' \in \mathcal{C}} s(a')}{\max_{a' \in \mathcal{C}} s(a') - \min_{a' \in \mathcal{C}} s(a')}, \tag{3}$$

and the fused score is the convex combination

$$\text{score}(a) = \alpha\, \hat{s}_{\text{text}}(a) + (1 - \alpha)\, \hat{s}_{\text{img}}(a), \qquad \alpha = 0.75. \tag{4}$$

The top-k articles by Eq. (4) form the evidence block. Three properties of this construction are load-bearing. First, the fused score only selects which articles enter the prompt; the generator sees headlines, passages, captions, and (for M_{vision}) image pixels, never scores or ranks. Second, $\alpha = 1$ eliminates the image term and recovers B_1 exactly along the identical code path: we verify this empirically, obtaining identical rankings on all 150 queries at $\alpha = 1$, so a B_1-vs-M difference cannot be an implementation artifact. Third, restriction to $\mathcal{C}$ precedes normalisation: normalising first would let excluded articles set the range and silently change what α means.

4.2 Generation and Abstention

All three systems share one `generate(prompt)` call to `gpt-4o-mini` at temperature 0 with a fixed seed, so any output difference is attributable to the evidence block, not sampling noise or a different model. Before generation, a retrieval-confidence gate checks $\max_{i \le k} s_{\text{text}}(a_i)$, the *raw*, un-normalized cosine (the fused score in Eq. (4) is min–max normalized per query and so is always ≈ 1 at its maximum regardless of evidence quality), against $\tau = 0.35$; below it, the system returns `INSUFFICIENT_EVIDENCE` without an LLM call. The model may also decline in free text without emitting that token, so any refusal rate is read off the generated text, not a single boolean flag.

4.3 Evaluation Protocol

Faithfulness is scored by two LLM-judge passes in the style of claim-decomposition factuality metrics [5,6]: (1) *decompose* a summary into atomic, self-contained claims, blind to the evidence, so claim boundaries are not shaped to fit it; (2) *verify* each claim against the evidence it was actually generated from, judged `supported` / `unsupported` by a model from a different family than the generator and blind to which system produced the summary. All comparisons are *paired* on items usable in both arms, reported with a 10,000-resample bootstrap 95% CI and a Wilcoxon signed-rank test; the frozen configuration declares which comparisons are prespecified before the final-test role is scored, so any other comparison is labelled post hoc rather than presented as confirmatory.

5 Numerical Experiments

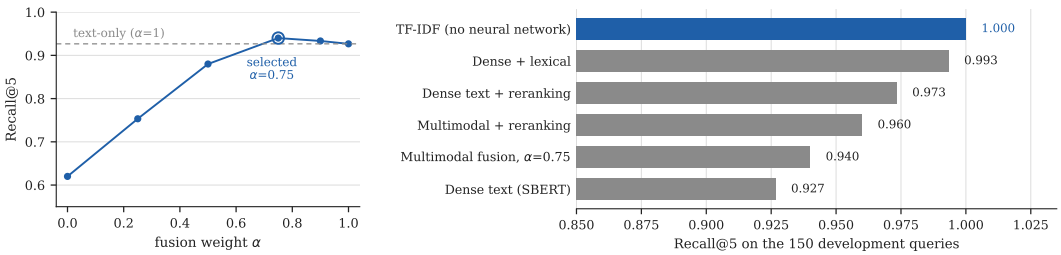
We build the system ladder of Section 4.1 and score it on the frozen 150-article final-test role. The protocol was run twice: first as a 20-article category-balanced pilot (12 paired usable cases), then on the complete role (the headline result reported below); nothing was re-selected between runs. All figures and statistics are computed by `python -m src.figures` directly from tracked experiment CSVs, so no number here can drift from the run that produced it.

5.1 Retrieval

On the frozen 20-query final sample (untouched during development), dense text, the selected fusion, and lexical retrieval all tie at Recall@5 $= 1.00$; image-only reaches only 0.55. Fusion therefore *ties* rather than improves on text retrieval once evaluated off the tuning set, consistent with Figure 1b: the query set carries enough lexical signal that little headroom remains for an image stream to change what gets retrieved, with or without reranking.

5.2 Faithfulness

Table 2 reports macro claim faithfulness on the full 150-article final test. B_0 receives no evidence and so never abstains (150/150 usable); its claims are judged against the union of B_1's and M's



(a) Fusion weight selection.

(b) Retrieval methods compared.

Figure 1: Retrieval on 150 development queries. (a) Recall@5 peaks at $\alpha = 0.75$ (selected), above the α=1 text-only reference; the inherited default $\alpha = 0.5$ was on the wrong side of the peak. (b) A parameter-free TF-IDF baseline reaches 1.00; reranking narrows the gap for both dense text ($0.927 \rightarrow 0.973$) and multimodal fusion ($0.940 \rightarrow 0.960$), but reranked dense text still beats reranked fusion: once lexical signal is available to both, the image stream does not help.

Table 2: Faithfulness on the frozen 150-article final test, usable cut.

System	Evidence given to the generator	Faithfulness	Usable
B_0	none (no retrieval)	0.146	150/150
B_1	text passages	0.864	77/150
M	fused retrieval + captions	0.874	80/150
M_{vision}	M + retrieved image pixels	**0.898**	84/150

retrieved evidence for the same query, with the same generator and judge as the rest of the ladder, so it is now a paired addition to this role rather than a number from a different pipeline. B_1, M, and M_{vision} instead gate on retrieval confidence and abstain roughly half the time (usable-cut refusal: 49% B_1, 47% M, 44% M_{vision}), so their usable counts are not directly comparable to B_0's. Every rung of the ladder improves the mean, and the single largest jump is $B_1 - B_0$: $+0.586$, 95% CI $[0.525, 0.643]$ over $n = 77$ paired usable items (bootstrap). Since the three retrieval systems are scored on overlapping but non-identical usable sets, Figure 2 instead reports the *paired* difference among them with a bootstrap CI and a Wilcoxon p-value (full numeric table in Appendix 7).

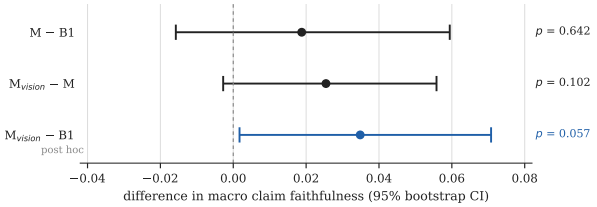


Figure 2: Paired differences in macro claim faithfulness, usable cut (95% bootstrap CI, Wilcoxon p). Both prespecified rungs cross zero; the post hoc full jump $M_{\text{vision}} - B_1$ is borderline ($p = 0.057$) but does not survive a Bonferroni correction for six comparisons ($p < 0.008$).

The frozen configuration declares the two adjacent rungs primary; neither reaches significance ($p = 0.642$, $p = 0.102$). The full jump $M_{\text{vision}} - B_1$, examined only after both prespecified tests were null, sits at the edge of significance but is post hoc and underpowered: confirming the smaller step alone would need roughly 224 paired items, well beyond this test role (Appendix 7). What is robust: every comparison is positive, and the two channel steps are each roughly half of the total accumulated effect.

5.3 Sample-Size Sensitivity

$M - B_1$ flips sign entirely between the two runs; the 20-article pilot also over/under-weighted section categories relative to the full role, so we treat it as superseded rather than an independent replication. The larger role improved the detectable-effect floor by 1.3–2.2$\times$ (Appendix 7), at the inference cost reported in Appendix 7.

5.4 Mechanism and Robustness

The population-level null cannot say whether images ever help or hurt. On evidence curated to be visually informative, $M_{\text{vision}} - M$ is $+0.205$ $[+0.057, +0.427]$ even though $M - B_1$ remains null on

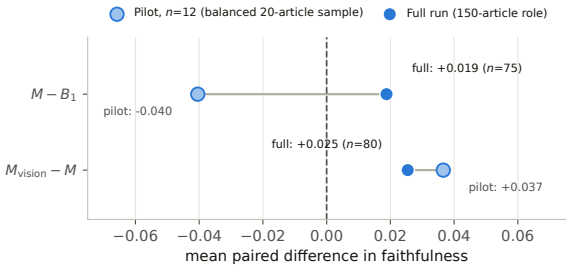


Figure 3: The identical frozen protocol at two sample sizes, 95% CI shown. At n=12 (grey, the pilot) fusion appeared to *reduce* faithfulness; at the full n=75–80 (blue, the full run) the sign reverses: the small-sample result was noise, not a finding.

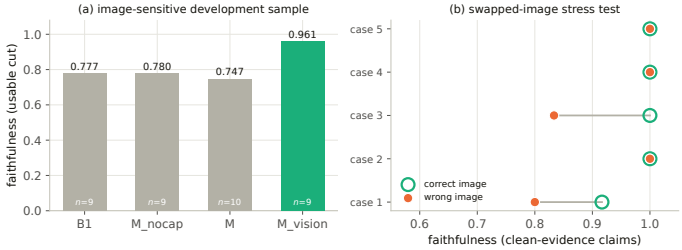


Figure 4: (a) On a 12-item sample curated for visual salience, M_{vision} reaches 0.961 vs. 0.75–0.78 for caption/text-only systems. (b) Swapping in another case's image on 5 cases with genuine pixel support drops faithfulness from 0.983 to 0.927. Pixels measurably help under salience and hurt under conflict, even though the population-average effect (Figure 2) is null.

the same items: the gain is specific to pixels. Swapping in an incorrect image on cases with genuine pixel support measurably reduces faithfulness, confirming the generator attends to image content. Only 7.1% of supported M_{vision} claims on the final-test role needed both text and pixel support, and none needed pixels alone: a small, conditional effect that dilutes to a population null at this sample size.

6 Discussion

Retrieval of any kind reduces hallucination dramatically: faithfulness rises from roughly 0.146 with no retrieval to 0.86–0.90 with any retrieval-augmented system, by far the largest effect in the study and unrelated to modality. Whether an image stream and captions help beyond text is null at the prespecified test ($M - B_1$, $p = 0.642$); Section 5.1 suggests why, since a zero-parameter TF-IDF baseline nearly saturates Recall@5, leaving little room for images to change what is retrieved. Whether actual pixels help beyond captions is also null at the prespecified comparison ($p = 0.102$), though directionally positive and borderline for the full text-to-pixel jump examined post hoc. The population average cannot say under which conditions images help or hurt, but the mechanism experiments (Figure 4) can: pixels measurably help under visual salience and hurt under image-text conflict, a real, small, conditional effect that this corpus does not create enough headroom to reveal on average, rather than pixels being inert.

7 Conclusions

Retrieval is the dominant lever against hallucination in this study; multimodal fusion and pixel-level grounding are directionally helpful but not statistically established at $n = 150$, and confirming even the smaller observed step would need roughly 224 paired items. Two things would most change the headline result: a larger paired test set (Appendix 7), and a generator that reasons jointly over multiple images rather than one caption at a time. Query design matters more than modality here: this query style is close to lexically saturated, so future work should choose queries with weaker lexical overlap to their gold evidence.

References

[1] Lewis, P., Perez, E., Piktus, A., et al. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

[2] Radford, A., Kim, J.W., Hallacy, C., et al. (2021) Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*.

[3] Reimers, N. & Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP 2019*.

[4] Johnson, J., Douze, M., & Jégou, H. (2019) Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

[5] Min, S., Krishna, K., Lyu, X., et al. (2023) FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *Proceedings of EMNLP 2023*.

[6] Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2023) RAGAS: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217*.

[7] OpenAI (2024) GPT-4o System Card. `https://openai.com/index/gpt-4o-system-card/`.

[8] RealTimeData. *bbc_news_alltime* (Hugging Face Datasets), config `2024-01`.

Appendix

A.1 Reproducibility

Corpus and splits are committed under `data/processed/` (not rebuilt: rebuilding is non-deterministic across machines). The frozen final configuration is `configs/final_research.json`. Figures and paired statistics regenerate with `python -m src.figures`, reading only tracked CSV/JSON under `results/experiments/`. Scaling the paired test set is bounded by the corpus (1,023 articles, 150 held out for final test); the retrieval-recall curve in Section 5.1 implies a substantially larger corpus would mainly add distractor articles, not change the multimodal comparison, so a bigger *test* role, not a bigger index, is what closing the 224-pair power gap requires.

A.2 Extended significance results

Three qualifications travel with the faithfulness results in Section 5.2. **Multiplicity:** six comparisons are reported across two usability cuts, uncorrected; a Bonferroni threshold would require $p < 0.008$, which nothing reaches. **Bootstrap vs. Wilcoxon:** the two tests agree in sign but disagree at the margin between usability cuts, the signature of a borderline effect. **Power:** at 80% power the minimum detectable effect is 0.043–0.054 faithfulness points against observed step sizes of 0.005–0.035.

Table 3: Paired differences, both usability cuts (Figure 2 plots the usable cut). Status records whether the comparison was declared in the frozen configuration before the final run; *nonhard* adds soft refusals to the usable cut.

Comparison	Status	n	Difference	95% CI	p
$M - B_1$	prespecified	75	$+0.019$	$[-0.016, +0.059]$	0.642
$M_{\text{vision}} - M$	prespecified	80	$+0.025$	$[-0.003, +0.056]$	0.102
$M_{\text{vision}} - B_1$	post hoc	75	$+0.035$	$[+0.002, +0.072]$	0.057
$M - B_1$ (nonhard)	prespecified	101	$+0.017$	$[-0.013, +0.051]$	0.706
$M_{\text{vision}} - M$ (nonhard)	prespecified	104	$+0.005$	$[-0.025, +0.036]$	0.520
$M_{\text{vision}} - B_1$ (nonhard)	post hoc	98	$+0.033$	$[-0.000, +0.067]$	0.043

Table 4: Cost of the 20-article pilot relative to the full 150-article role, usable cut. Mean paired difference in faithfulness; the pilot's 95% CI is wide enough to contain both zero and the full run's estimate.

Comparison	Pilot (n=12)	Full run (n=75–80)
$M - B_1$	-0.040	$+0.019$
$M_{\text{vision}} - M$	$+0.037$	$+0.025$

A.3 Efficiency and Cost

Embedding takes about 40 s on two CPU threads; median retrieval is 35 ms/query. The full research extension cost $0.92 across 1,008 hosted-inference calls against a $8 software-budget ceiling: $0.71 of that was the 150-article rerun of the frozen protocol (Section 5.3), and a further $0.18 across 450 calls came from adding the B_0 arm to that same role (Table 2), for a running total of $1.10 across 1,458 calls. All spend stayed within the ceiling.