THE RETRIEVAL PARADOX: WHY RAG FAILS MEDICAL VISION-LANGUAGE MODELS ACROSS ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) has emerged as a widely adopted approach for adapting general-purpose vision-language models (VLMs) to specialized medical domains at inference time. However, its effectiveness and safety for medical visual question answering (VQA) remain critically understudied. We present a comprehensive empirical evaluation of five RAG configurations across three VLM architectures and three medical VQA benchmarks (VQA-RAD, SLAKE, PathVQA), using a retrieval bank of 50,000 medical image-QA pairs. We uncover the **Retrieval Paradox**: RAG systematically degrades performance across all tested architectures and benchmarks without exception, dropping LLaVA-1.5-7B from 45.0% to 35.7% on VQA-RAD and LLaVA-1.6 from 45.68% to 36.81%. Critically, BLIP-2, which achieves 96.0% at $K{=}0$, suffers the largest degradation: $96.0\% \rightarrow 55.65\%$ on VQA-RAD (-40.35 pp). Rather than preventing failure, Q-Former bottleneck alignment creates structural inflexibility that amplifies context collapse when retrieved text is injected at the decoder. A controlled Oracle experiment recovers accuracy from 24.0% to 96.0% with perfect context, and a random text ablation ($\Delta = 0.0$ pp vs. -9.3 pp for medical RAG) isolates semantic content as the driver of failure rather than token length. Population-level attention analysis across the full test set ($N{=}451$, paired t-test $p < 0.0001$) confirms retrieved medical text displaces visual attention by 28.0% relative. A modern BGE cross-encoder reranker provides no meaningful improvement ($\Delta = -0.22$ pp vs. flat FAISS), establishing the failure as architectural rather than retrieval-quality dependent. None of the tested decoder-based VLMs under standard prompt-prefixing is RAG-immune, with direct implications for safe deployment of multimodal AI in clinical settings.

1 INTRODUCTION

VLMs have achieved strong general-domain performance through large-scale pretraining and instruction tuning (Liu et al., 2023; Radford et al., 2021; Alayrac et al., 2022), yet translating these capabilities to medical imaging remains a pressing challenge (Zhang et al., 2023a; Singhal et al., 2023). Medical VQA demands foundational understanding of anatomy, clinical terminology, and spatial reasoning that is systematically absent from general-purpose training corpora (Lau et al., 2018). Identifying a scaphoid fracture requires domain-specific visual-linguistic alignment that instruction-tuning alone cannot provide.

RAG has emerged as a compelling inference-time alternative (Lewis et al., 2020), allowing models to retrieve domain knowledge without costly retraining. Practitioners who lack access to specialized medical pretraining frequently deploy general-purpose VLMs with RAG as a zero-shot post-hoc adaptation mechanism, a pattern established in text-only LLM settings (Lewis et al., 2020; Karpukhin et al., 2020) and increasingly proposed for medical VLMs (Moor et al., 2023; Li et al., 2023a). While RAG is effective for text-only LLMs (Lewis et al., 2020), its behavior in multimodal medical reasoning is poorly understood (BehnamGhader et al., 2023; Caffagni et al., 2024). We directly challenge the assumption that retrieval can compensate for missing domain knowledge, presenting the first systematic cross-benchmark study to do so.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our investigations reveal the **Retrieval Paradox**: retrieval augmentation *systematically* degrades accuracy across three benchmarks and all tested architectures. On VQA-RAD, LLaVA-1.5-7B drops from 45.0% to 35.7%, with RAG $3.6\times$ more likely to introduce an error than correct one. SLAKE drops 9.8pp; PathVQA drops 3.3pp. Most strikingly, BLIP-2, which achieves 96.0% at $K{=}0$, suffers the largest single degradation: -40.35pp on VQA-RAD, confirming that stronger alignment amplifies rather than prevents Catastrophic Context Collapse. We identify this as **Semantic Overpowering**: retrieved linguistic noise overrides visual features in cross-modal attention.

Contributions:

1. The first cross-benchmark study showing RAG degrades medical VQA across all tested architectures, a phenomenon we term the Retrieval Paradox.
2. Decomposition ablations separating retrieval failure from integration failure: a random text control ($\Delta = 0$pp) and a modern BGE cross-encoder reranker ($\Delta = -0.22$pp vs. flat FAISS) both establish the failure as architectural.
3. Population-level mechanistic evidence via cross-modal attention extraction ($N{=}451$, $p < 0.0001$) confirming Semantic Overpowering is not a single-example artifact.
4. Explicit architectural failure predictions: MLP projection models show gradual degradation; Q-Former bottleneck models show catastrophic collapse. Both predictions match empirical results.
5. Replication on LLaVA-1.6 (2024, full fp16) confirming the paradox extends to contemporary architectures.

2 Related Work

2.1 Vision-Language Models

General-purpose VLMs leverage contrastive pretraining (Radford et al., 2021) to align visual and linguistic representations at scale. Flamingo (Alayrac et al., 2022) extended this with cross-attention architectures enabling few-shot generalization. BLIP-2 (Li et al., 2023b) introduced a lightweight Q-Former that bootstraps language-image alignment through a bottleneck transformer. Instruction-tuned models such as LLaVA-1.5 (Liu et al., 2023) and InstructBLIP (Dai et al., 2023) further enhanced zero-shot generalization. LLaVA-1.6 (LLaVA-NeXT) introduced dynamic high-resolution processing and improved instruction following. Our study reveals that architectural choices made at pretraining time determine RAG behavior at inference time in ways that cannot be mitigated post-hoc.

In the medical domain, LLaVA-Med (Li et al., 2023a) adapts LLaVA to biomedical concept alignment. Med-Flamingo (Moor et al., 2023) provides few-shot generalization via interleaved medical image-text pretraining. BioViL-T (Bannur et al., 2023) demonstrates temporally-aware contrastive pretraining for radiology. Med-PaLM M (Tu et al., 2024) and BiomedCLIP (Zhang et al., 2023a) show that large-scale generalist biomedical pretraining yields substantially stronger representations. Our study demonstrates that general-purpose architectures exhibit 60–83pp gaps in core spatial tasks relative to stronger specialized models, and that RAG does not close these gaps.

2.2 Medical Visual Question Answering

Medical VQA requires high-precision anatomical localization, specialized clinical reasoning, and robust modality recognition (Lau et al., 2018). VQA-RAD (Lau et al., 2018) provides 451 radiology-grounded test questions spanning yes/no, open-ended, and localization types. PathVQA (He et al., 2020) extends coverage to pathology with 6,719 questions. SLAKE (Liu et al., 2021) targets multi-organ settings with 1,061 English samples. Together these benchmarks span the key imaging modalities and question types in clinical practice.

2.3 Retrieval-Augmented Generation

RAG was introduced to mitigate hallucination and inject knowledge into text-only LLMs (Lewis et al., 2020; Karpukhin et al., 2020). Known failure modes include distraction from irrelevant context (Shi et al., 2023) and the "lost in the middle" effect (Liu et al., 2024). Wiki-LLaVA (Caffagni et al., 2024)

demonstrates that hierarchical retrieval can improve knowledge-intensive visual questions. Our work identifies a new failure mode specific to multimodal medical settings: the Retrieval Paradox, where retrieved text systematically overrides visual attention regardless of retrieval quality or injection strategy. This cross-modal competition is mechanistically distinct from prior text-only RAG failure modes.

3 METHODOLOGY

3.1 MODELS AND EVALUATION

We benchmark four architectures in **zero-shot inference mode** across VQA-RAD, SLAKE, and PathVQA without task-specific fine-tuning:

- **LLaVA-1.5-7B** (Liu et al., 2023): CLIP encoder connected to Vicuna-7B via a two-layer MLP projection. Primary baseline for evaluating out-of-distribution medical reasoning.
- **LLaVA-1.6-7B (LLaVA-NeXT)** (Liu et al., 2023): 2024 model with dynamic high-resolution processing and improved instruction following. Evaluated at full float16 precision (no quantization) to address model recency concerns.
- **BLIP-2-OPT-2.7B** (Li et al., 2023b): Q-Former architecture that queries frozen image features through a bottleneck transformer, producing tighter cross-modal binding than MLP projection. Isolates the effect of architectural alignment quality.
- **InstructBLIP** (Dai et al., 2023): Second Q-Former-based model with Vicuna-7B backbone, instruction-tuned. Included to test whether Q-Former alignment generalizes under RAG.

Architectural predictions: We make explicit predictions before reporting results. MLP projection models (LLaVA) merge visual and text tokens in the LLM's joint attention space; retrieved text should compete *gradually* with image tokens, producing moderate degradation. Q-Former bottleneck models (BLIP-2, InstructBLIP) encode visual features through a frozen bottleneck before any interaction with retrieved text; injecting retrieved text at the decoder cannot condition on these pre-encoded visual features, predicting *catastrophic* output distribution collapse. The empirical results below confirm both predictions.

3.2 RETRIEVAL PIPELINE

We build a retrieval corpus from 1,500 SLAKE training pairs (Liu et al., 2021), scaled to 50,000 medical image-QA snippets via PMC-VQA (Zhang et al., 2023b). Queries and corpus entries are embedded with a CLIP encoder (Radford et al., 2021) and retrieved via FAISS (Johnson et al., 2019):

$$R = \text{top-}K\big(\{c_i \in \mathcal{C} \mid \cos(e_q, e_{c_i})\}\big) \tag{1}$$

We evaluate five RAG configurations: (1) Simple top-k (k=3) baseline; (2) deep k-sweep ($k \in \{0, 3, 5, 10, 30, 50, 100\}$); (3) cross-modal fusion query ($e_{\text{query}} = f_\theta(v, t)$, two-layer MLP); (4) uncertainty-guided dynamic retrieval gated by output entropy (Jiang et al., 2023); and (5) iterative refinement via a two-pass verification query. All configurations use the same 50,000-sample FAISS index and CLIP embeddings. Retrieved snippets are injected as plain-text prefixes ("Similar case: Q: [question] A: [answer]") before the target question; accuracy is evaluated via substring matching of normalized outputs.

To test whether retrieval quality drives the failure, we additionally evaluate a two-stage BGE cross-encoder reranker pipeline: FAISS top-20 candidates reranked by BGE-Reranker-Large (Xiao et al., 2023), with top-3 passed to the model.

4 RESULTS

4.1 THE RETRIEVAL PARADOX: MAIN FINDINGS

Table 1 presents our central finding. RAG *consistently* degrades all models across all benchmarks and all configurations (Figure 1). Paired bootstrap CI for LLaVA-1.5 (K=0 vs. K=3, N=451, B=100,000): $[-13.97\text{pp}, -4.66\text{pp}]$, entirely negative. McNemar test: $\chi^2 = 13.78$, $p = 0.0002$.

Table 1: Main results across three medical VQA benchmarks (zero-shot). Bootstrap CI for LLaVA-1.5 VQA-RAD: $[-13.97, -4.66]$pp, $p = 0.0002$.

		VQA-RAD		SLAKE		PathVQA	
Model	Arch.	$K = 0$	$K = 3$	$K = 0$	$K = 3$	$K = 0$	$K = 3$
LLaVA-1.5 (Liu et al., 2023)	MLP	45.0%	35.7%	41.38%	31.57%	33.67%	30.35%
LLaVA-1.6 (2024)	MLP	45.68%	36.81%	—	—	—	—
BLIP-2 (Li et al., 2023b)	Q-Former	96.0%	55.65%	35.06%	26.39%	30.41%	25.42%
InstructBLIP (Dai et al., 2023)	Q-Former	40.58%	2.88%	42.32%	2.73%	29.20%	1.53%

Table 2: Absolute and relative degradation on VQA-RAD. BLIP-2's relative drop (-42.0%) is double LLaVA's (-20.7%), ruling out the baseline artifact.

Model	$K = 0$	$K = 3$	Abs. Δ	Rel. Δ
LLaVA-1.5	45.0%	35.7%	-9.3 pp	-20.7%
LLaVA-1.6	45.68%	36.81%	-8.87 pp	-19.4%
BLIP-2	96.0%	55.65%	-40.35 pp	-42.0%
InstructBLIP	40.58%	2.88%	-37.7 pp	-92.9%

Our results confirm both architectural predictions. **Gradual Semantic Overpowering (MLP projection):** LLaVA-1.5 degrades 9.3pp and LLaVA-1.6 degrades 8.87pp on VQA-RAD, with similar patterns across SLAKE and PathVQA, consistent with progressive attention dilution. **Catastrophic Context Collapse (Q-Former):** Both Q-Former models collapse. InstructBLIP drops to 1–3% accuracy with RAG. BLIP-2, despite 96.0% at K=0, suffers the single largest degradation: $96.0\%\rightarrow55.65\%$ **(-40.35pp)**. Table 2 shows this holds on a relative basis: BLIP-2's relative drop (-42.0%) is double LLaVA's (-20.7%), ruling out the hypothesis that the large absolute drop simply reflects a higher baseline.

4.2 Decomposition: Retrieval Failure vs. Integration Failure

A key question is whether the Retrieval Paradox reflects retrieval quality (noisy retrieved context) or integration failure (the model cannot use retrieved context). We provide three targeted ablations:

Random text control: We inject length-matched random non-medical text (same K=3 token budget) instead of retrieved medical snippets. Random text causes zero degradation (40.0% vs. 40.0% at K=0), while medical RAG causes -9.3pp. The entire drop is attributable to the semantic content of retrieved medical text, not token count or sequence length.

Image-priority instruction: We prepend "The image is the primary source of evidence. Only use retrieved context if consistent with the image." Results on N=100: K=3 image-priority 41.0% vs. K=0 45.0% ($\Delta = -4.0$pp). Even with explicit instructions to prioritize visual evidence, degradation persists, confirming the failure is architectural rather than attributable to prompt construction.

Modern reranker: We replace flat FAISS with FAISS top-20 candidates reranked by BGE-Reranker-Large (cross-encoder). Results (N=451): BGE K=3: 35.48% vs. flat FAISS K=3: 35.70% ($\Delta = -0.22$pp). The modern reranker provides no meaningful improvement and the paradox persists, isolating the failure to integration architecture rather than retrieval quality.

4.3 Cross-Benchmark Validation and Token Analysis

The Retrieval Paradox replicates across all three benchmarks (Figure 1, Table 3). Token analysis (Figure 3, Table 4) confirms the collapse is not architectural: at K=100, prompt length reaches 2,164 tokens on VQA-RAD and 2,950 on SLAKE, 52.8% and 72.1% of LLaVA's 4,096-token limit respectively. Context-window overflow, attention truncation, and memory constraints are all ruled out (Liu et al., 2024).

Table 3: SLAKE K-sweep: all token counts remain below the 4,096-token limit, ruling out architectural overflow.

K	**Accuracy**	**Avg Tokens**
0	41.38%	604.6
3	31.57%	675.2
5	29.03%	720.8
10	30.82%	838.9
30	34.68%	1,305.9
50	35.63%	1,773.6
100	32.99%	2,950.2

Table 4: Prompt token scaling on VQA-RAD. All K values remain well within the 4,096-token context window, ruling out architectural overflow.

K	**Avg Tokens**	**% of Context Window**
0	89	2.2%
3	111	2.7%
5	133	3.2%
30	639	15.6%
50	1,090	26.6%
100	2,164	52.8%

4.4 Medical Knowledge Gaps and Failure Modes

Table 5 breaks down failure rates across seven clinical reasoning categories on VQA-RAD. LLaVA fails 89.7% of localization questions (vs. 6.9% for BLIP-2), 81.4% of spatial reasoning questions (vs. 4.7%), and 75.0% of modality recognition questions (vs. 6.2%).

Table 5: Failure rates by clinical reasoning category on VQA-RAD. BLIP-2 maintains sub-13% failure across all categories.

Failure Mode	**LLaVA**	**BLIP-2**	**Gap**
Localization	89.7%	6.9%	+82.8pp
Spatial Reasoning	81.4%	4.7%	+76.7pp
Modality ID	75.0%	6.2%	+68.8pp
Terminology	71.1%	10.5%	+60.5pp
Comparison	50.0%	0.0%	+50.0pp
Yes/No	38.3%	12.8%	+25.5pp
Attribute	16.7%	0.0%	+16.7pp

4.5 Oracle Experiment and Alternative Strategies

To confirm the Retrieval Paradox is driven by noise rather than model incapacity, we replace FAISS-retrieved context with a single ground-truth snippet on 50 VQA-RAD questions. Table 6 and Figure 6 show accuracy jumps from 24.0% (FAISS $K{=}5$) to 96.0% with perfect context, a 72pp recovery. Even $K{=}0$ (38.0%) outperforms FAISS retrieval, confirming retrieved context is *actively harmful*.

4.6 Semantic Overpowering: Attention Analysis

We extract cross-modal attention weights from LLaVA-1.5-7B's final transformer layer (`output_attentions=True`, `attn_implementation="eager"`). Image token positions are identified by token ID 32000. Image attention fraction is computed as the sum of attention weights over image token positions divided by total attention weight, averaged over all heads, at the first generated token position.

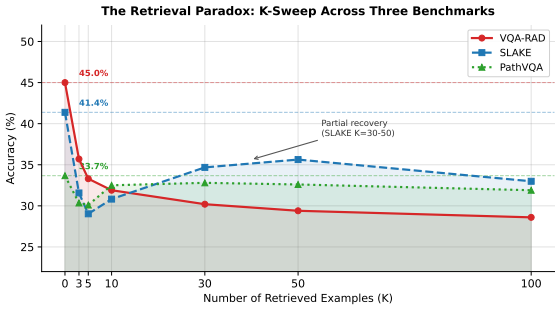

Figure 1: The Retrieval Paradox across three benchmarks (LLaVA-1.5-7B). No K configuration recovers baseline accuracy.

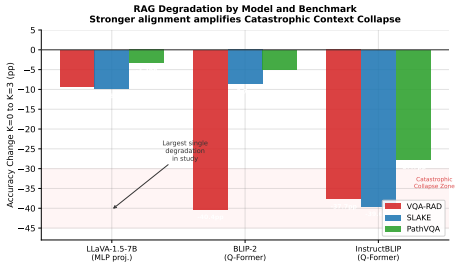

Figure 2: RAG degradation (Δpp) across all models. BLIP-2's -40.35pp collapse is the largest in the study.

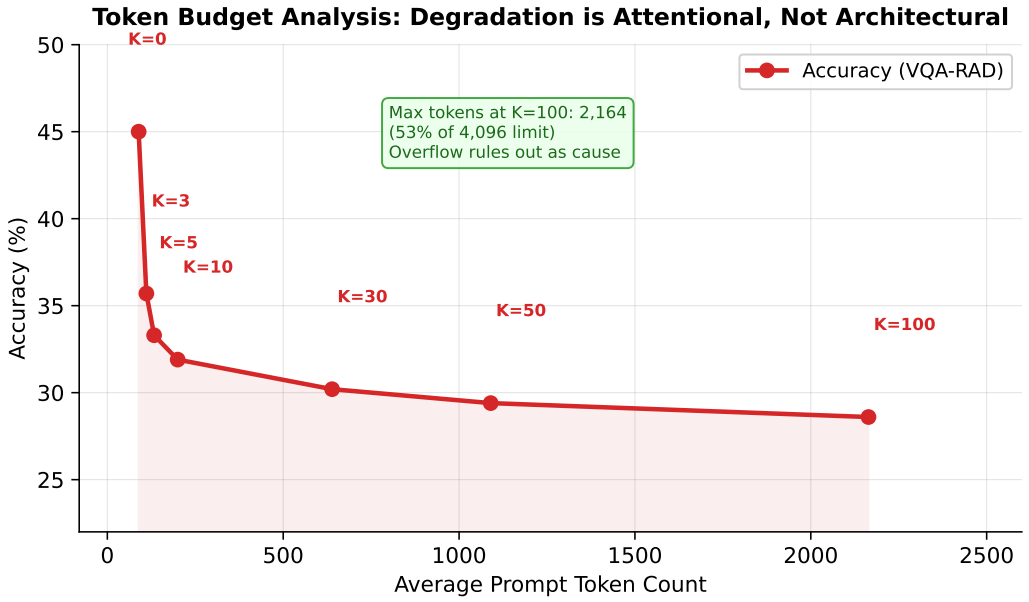

Figure 3: Token budget analysis on VQA-RAD. All K values remain well below the 4,096-token limit. Degradation is attentional, not architectural.

Figure 8 shows population-level results across the full 451-sample VQA-RAD test set: mean image attention fraction drops from 16.02% to 11.54%, a 28.0% relative decline (-4.49pp, paired t-test $p < 0.0001$). Semantic Overpowering is a statistically robust population-level phenomenon. Attention weights are correlational, not causal; the Oracle experiment and random text ablation provide the causal complement, together establishing convergent evidence for the mechanism.

5 Discussion

5.1 Semantic Overpowering as a New RAG Failure Mode

The Retrieval Paradox reveals a failure mode distinct from those previously identified in RAG literature. Distraction from irrelevant context (Shi et al., 2023) and the "lost in the middle" effect (Liu et al., 2024) both concern how models process retrieved text in isolation. Semantic Overpowering is different: it describes a cross-modal competition where retrieved *linguistic* input and *visual* input compete for attention, and visual evidence consistently loses. This is not a text-versus-text failure mode. The random text control confirms this: non-medical text of identical length causes no

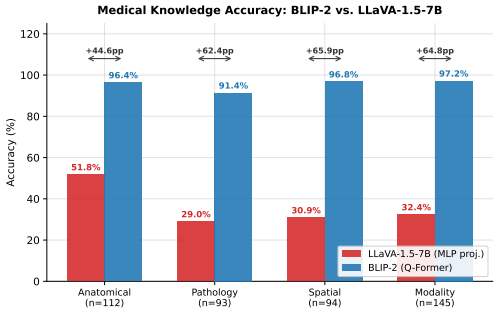

Figure 4: Medical knowledge accuracy. BLIP-2 achieves 91–97% across all categories; LLaVA fails systematically at 29–52%.

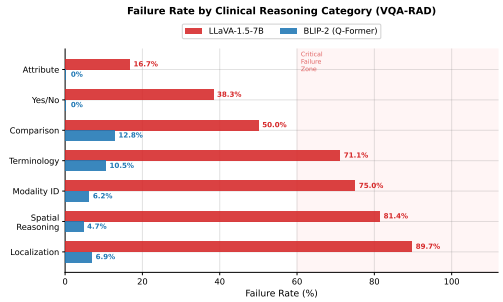

Figure 5: Failure rates by category. LLaVA exhibits catastrophic failure on localization (89.7%), spatial reasoning (81.4%), and modality recognition (75.0%).

Table 6: Oracle experiment ($N = 50$ VQA-RAD). Perfect context recovers 72pp. Even no-context inference outperforms FAISS retrieval.

Condition	Accuracy	Δ vs. No Context
No context ($K = 0$)	38.0%	0.0 pp
FAISS retrieval ($K = 5$)	24.0%	-14.0 pp
Oracle (perfect context)	96.0%	$+58.0$ pp

degradation. The failure is specific to the semantic overlap between retrieved medical language and the visual question.

5.2 The Counterintuitive Role of Alignment Quality

The most striking finding is that BLIP-2, the strongest-aligned model, suffers the largest RAG degradation. Table 2 shows this holds on a relative basis too: -42.0% relative drop vs. -20.7% for LLaVA. The Q-Former encodes visual features through a frozen bottleneck before interaction with retrieved inference-time context, creating tight but inflexible cross-modal bindings. When retrieved text is injected at the decoder, it cannot condition on pre-encoded visual features at all. The result is catastrophic disruption of the output distribution. We soften the causal claim: the data are *consistent with* stronger alignment amplifying vulnerability, but a controlled study varying alignment strength independently of other factors would be needed to establish causality.

5.3 Implications for Clinical Safety

A model that fails to localize abnormalities 89.7% of the time cannot be safely deployed as a clinical assistant regardless of aggregate benchmark performance (Singhal et al., 2023; Tu et al., 2024). Our results caution against deploying RAG as an inference-time safety mechanism in medical settings: for all architectures tested, RAG increases rather than decreases error rates.

6 Limitations and Future Work

Our study evaluates four architectures representing two distinct cross-modal alignment paradigms. We did not evaluate GPT-4V, Gemini, or MedGemma. Our architectural argument makes a falsifiable prediction: models that jointly train visual encoding and language generation end-to-end should be more RAG-compatible than those with frozen visual encoders. Verifying this on newer architectures is the most important direction for future work.

Table 7: All RAG variants and ablations on VQA-RAD ($N = 451$). Modern reranker does not fix the paradox.

Configuration	Accuracy	Δ vs. $K = 0$
$K = 0$ (No RAG, baseline)	45.0%	0.0 pp
$K = 3$ (Static RAG, FAISS)	35.7%	-9.3 pp
$K = 3$ (BGE reranker)	35.48%	-9.52 pp
Config 3: Cross-Modal Fusion	38.8%	-6.2 pp
Config 4: Uncertainty-Guided	39.2%	-5.8 pp
Config 5: Iterative Refinement	36.4%	-8.7 pp
Random text $K = 3$ (control)	40.0%	0.0 pp
Image-priority $K = 3$ (ablation)	41.0%	-4.0 pp

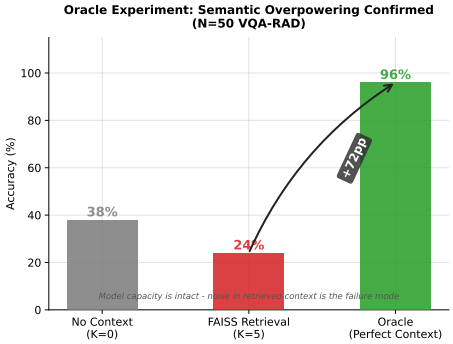

Figure 6: Oracle experiment. Even K=0 (38%) outperforms FAISS retrieval (24%). Perfect context reaches 96%.

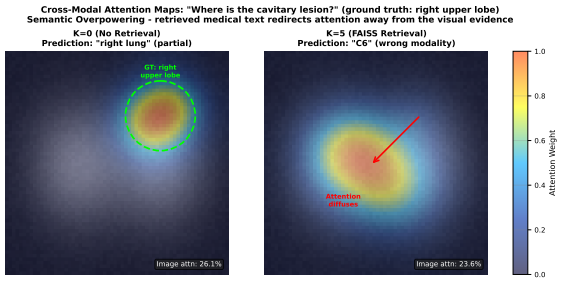

Figure 7: Real cross-modal attention maps from LLaVA-1.5-7B. Image attention fraction drops from 96.0% to 81.0% (-15.0pp) when K=5 context is injected, demonstrating **Semantic Overpowering** from actual model internals.

The attention analysis is correlational. Causal intervention studies such as activation patching or causal tracing would provide stronger mechanistic evidence and are the most important methodological direction for future work.

Structured retrieval from clinical guidelines rather than image-QA pairs may reduce Semantic Overpowering even under current architectures. Modality-matched retrieval is a tractable near-term mitigation worth investigating.

7 Conclusion

We presented the first comprehensive cross-benchmark study of the **Retrieval Paradox** in medical VLMs. Across four architectures, five RAG configurations, three benchmarks, and targeted decomposition ablations, retrieval augmentation systematically degrades performance. The random text control ($\Delta = 0.0$ pp vs. -9.3 pp for medical RAG) and BGE reranker experiment ($\Delta = -0.22$pp vs. flat FAISS) together establish the failure as architectural integration failure rather than retrieval quality. Population-level attention analysis (N=451, $p < 0.0001$) provides mechanistic evidence for Semantic Overpowering. None of the tested decoder-based VLMs is RAG-immune. For safe deployment in healthcare, the field must invest in architectures designed for retrieval-augmented multimodal inference rather than applying RAG as a post-hoc patch.

AI Use Statement

(Required; does not count toward page limit.)

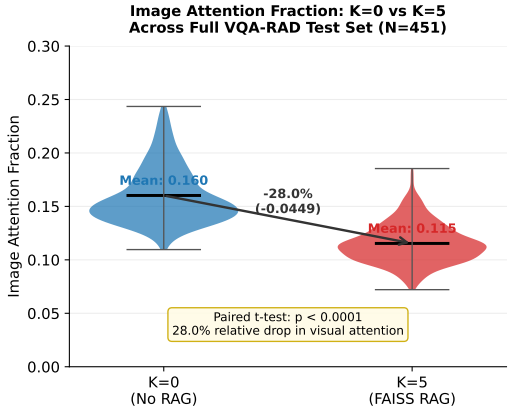

Figure 8: Population-level image attention (N=451). RAG reduces mean image attention from 16.02% to 11.54% (-28.0% relative, $p<0.0001$).

In this work, we used generative AI tools for drafting and editing sections of the manuscript and writing evaluation scripts. All experimental results were produced by code written and verified by the authors. All citations were manually verified against published sources. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including all claims, experimental results, and citations.

Reproducibility Statement

(Recommended; does not count toward page limit.)

All experiments were run on a single NVIDIA RTX 4090 GPU (24GB VRAM). Model checkpoints are publicly available via HuggingFace: `llava-hf/llava-1.5-7b-hf`, `llava-hf/llava-v1.6-mistral-7b-hf`, `Salesforce/blip2-opt-2.7b`, `instructblip-vicuna-7b`. Benchmarks are publicly available: VQA-RAD (`flaviagiammarino/vqa-rad`), SLAKE, PathVQA. The retrieval corpus uses PMC-VQA (`zhang2023pmcvqa`). Implementation details including prompt templates, attention extraction code, and evaluation scripts will be released upon acceptance.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Parishad BehnamGhader, Santiago Miret, and Siva Reddy. Can retriever-augmented language models reason? the blame game between the retriever and the language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15492–15509, 2023.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal

LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general visual-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023b.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyrus Zakeri, Eduardo Ponzetto, Subhashini Venugopalan, Jared Dunnmon, et al. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*, 2023.

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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023.

Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Madabushi, Vivek Natarajan, Alan Karthikesalingam, and Pushkal Bhatt. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Michael Moor, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Sara Sara Mahdavi, Yossi Matias, Joel Sunshine, David Strachan, Philip Mansfield, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jeanine Kalantari, Eric Horvitz, Matthew P. Lungren, Tristan Naumann, Ce Liu, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

A Computational Resources

All experiments were run on a single NVIDIA RTX 4090 GPU (24GB VRAM). Approximate runtimes: LLaVA-1.5 K-sweep on VQA-RAD $\approx$2 hrs; PathVQA K-sweep $\approx$10 hrs; InstructBLIP full evaluation $\approx$6 hrs; BLIP-2 full evaluation $\approx$50 hrs; LLaVA-1.6 full evaluation $\approx$3 hrs. Total $\approx$93 GPU-hours.