Separation Without Transfer: Contrastive Fine-Tuning Transfers to Held-Out Retrieval for Domain-Native Encoder in Mammography Report Generation

Duy Nguyen and Wenjing Yang†
Department of Mathematics, Seattle University,
Seattle, WA 98122, USA
E-mail: dnguyen44@seattleu.edu †*E-mail: wyang1@seattleu.edu*

Retrieval-augmented generation (RAG), where an encoder retrieves similar prior studies to condition a vision–language model (VLM), is increasingly applied in automated mammography reporting. Prior work fine-tunes the decoder and holds the retrieval encoder fixed, leaving open whether the performance is driven by the encoder's substrate, the retrieval representation, or their interaction. This work addresses the gap by isolating the retrieval encoder. Three VLM decoders (MedGemma, Qwen2.5-VL, and LLaVA-Med) remain frozen while four retrieval encoders, each pretrained on a different image domain, are evaluated before and after the same supervised-contrastive fine-tuning on a study-level VinDr-Mammo split. Experiments show that the three out-of-domain encoders retrieve *significantly below* the majority-class floor of P@1 = 0.487. The mammography-native encoder, Mammo-CLIP, alone increases above the floor ($0.513 \rightarrow 0.585$, $p = 3.3 \times 10^{-10}$). With the decoder frozen, performance gain reaches the report, increasing suspicion accuracy +20.7% and +15.6% on the two retrieval-sensitive decoders, and the same gain appears at the finer five-class BI-RADS resolution (+23.2% and +14.6%). Additional analysis of the gain generated based on the ACR management band assesses whether there is actual practical improvement in the clinical setting.

Keywords: Retrieval-Augmented Generation; Vision–Language Models; Mammography; Report Generation; Contrastive Learning; Representation Geometry.

1. Introduction

Automated radiological image-to-report generation remains an important field of research in the era of artificial intelligence in medical and healthcare applications.[1] Healthcare providers use mammograms, or mammography, to look for early signs of breast cancer before symptoms develop. A clinically useful system must preserve subtle image findings, distinguish healthy, benign, and suspicious studies, and produce structured outputs that support radiologist review. The task is inherently multimodal: a vision–language model (VLM) interprets a query mammogram while using text from prior reports to generate a structured assessment. Such systems could streamline screening and diagnostic decision support, but current VLMs that read a mammogram and draft structured reports still fall short of radiologist-level accuracy in practice.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Retrieval-augmented generation (RAG) grounds a generative model in retrieved evidence and has demonstrated practical value in knowledge-intensive generation.[2] In medical and healthcare applications, the pipeline includes an encoder that retrieves similar prior studies in terms of diagnosis or clinically relevant findings, selects the neighbors injected into the prompt, and a VLM decoder that reads the query image and retrieved text and writes the report. Encoder fine-tuning changes the neighborhood structure of the retrieval space and therefore changes the evidence available to every downstream decoder. Decoder fine-tuning may improve general task performance, but may not improve the incremental benefit attributable to RAG. The bottleneck is therefore modality-specific: it sits in the visual retrieval step, which has no text-side analog. Extending to multimodal RAG, the mammography-RAG pipeline[3] fine-tunes the VLM decoder through QLoRA,[4] while fixing retrieval to one frozen OpenCLIP encoder,[5] leaving unclear whether performance is driven by the encoder's substrate, the retrieval representation, or their interaction. This also creates a data-centric problem: before clinical validation, the encoder must preserve clinically relevant detail, retrieve appropriate evidence despite class imbalance, and generalize beyond its fine-tuning sample.

We address this gap through a controlled component-wise evaluation and retrieval-geometry analysis with supervised contrastive encoder fine-tuning. Three VLM decoders, including MedGemma, Qwen2.5-VL, and LLaVA-Med[6–8] remain frozen, while four retrieval encoders are compared, each pretrained on a different domain: generic web images, biomedical figures, organism imagery, and mammography-native data. Each encoder is evaluated before and after fine-tuning with the same supervised-contrastive (SupCon) objective.[9] On a study-level VinDr-Mammo split,[10] we evaluate two endpoints within structured report generation: the three-class suspicion label and the five-class BI-RADS category. Held-out nearest-neighbor precision, class-separation geometry, train-to-test geometry gaps, and within-decoder report agreement connect upstream retrieval behavior to downstream generation. The majority-class floor (P@1 = 0.487) exposes imbalance-related failure, while paired frozen-versus-fine-tuned comparisons distinguish improved training fit from transferable signal. Together, these measures provide practical tools for auditing, monitoring, and interpreting a multimodal healthcare RAG system rather than relying only on aggregate generation scores. The main contributions are as follows:

- Holding three VLM decoders fixed and varying only the retrieval encoder, we show that the benefit of supervised-contrastive fine-tuning is conditional on the encoder, transferring to held-out retrieval only for the mammography-native encoder.
- We identify input resolution—224×224 composite versus native 1520×912 views—as a plausible but not fully isolated driver of which encoders generalize under identical fine-tuning.
- Within-decoder contrasts localize the failure at the retrieval layer: retrieval-level generalization reappears in report agreement for Qwen2.5-VL and LLaVA-Med and not for MedGemma, which is insensitive to the strongest encoder's retrieval quality though not to BiomedCLIP's.
- We decompose the generated report gain by ACR management band (routine screening, short-interval follow-up, and biopsy) to measure the improvement under the clinical setting.

- We provide a component-wise auditing protocol for any retrieval-augmented VLM used in mammography report generation: held-out P@1 against the majority-class floor, the train/test margin ratio, and paired within-decoder tests under multiplicity control.

2. Related Work

RAG in medicine. RAG augments a generator with nonparametric memory retrieved from an external corpus.[2] Medical variants have emphasized literature retrieval, iterative querying, and benchmarking of retriever–generator combinations.[11–13] Lozano *et al.*[11] answer medical questions from the live scientific literature, and Xiong *et al.*[12] add iterative follow-up querying so a model refines what it retrieves across rounds. These systems improve retrieval and prompting over *text* corpora, where query and indexed evidence are already language and the encoder is an off-the-shelf text embedder. MedRAG/MIRAGE[13] makes the retriever the lever for text medical QA: with the generator fixed, all four off-the-shelf retrievers beat no-retrieval (60.7), but retriever choice alone spans 64.6 to 70.1, which is a gap the authors trace to SPECTER's training-objective mismatch. These studies show that retriever choice can materially change downstream accuracy even when the generator is fixed. Most medical RAG benchmarks are text-to-text: the query and the indexed evidence are already linguistic. Our setting instead requires a visual encoder to infer clinical similarity directly from mammograms.

Mammography report generation and medical VLMs. Automated image-to-report generation remains an open problem. Zhang *et al.*[1] characterize it as unsolved despite rapid progress in vision–language modeling. RAG is a leading proposal for closing the gap, and MammoWise[3] is a recent RAG-for-mammography pipeline that retrieves similar prior studies to condition a vision–language decoder. It combines local retrieval with MedGemma, Qwen2.5-VL, and LLaVA-Med for structured mammography reporting.[3,6–8] It fine-tunes the decoder and holds the retrieval encoder fixed, which is the design decision we invert. VinDr-Mammo provides four-view full-field digital mammography studies and radiologist annotations, but not biopsy-confirmed malignancy for every case.[10]

Contrastive representation learning. Contrastive image–text pretraining aligns paired modalities in a shared space. CLIP and OpenCLIP learn normalized image–text representations from large web corpora.[5,14] Supervised contrastive learning extends the objective to pull same-label examples together using annotations.[9] Wang and Isola[15] give a geometry-level account of what such objectives optimize, through alignment and uniformity on the hypersphere. A parallel line specializes the substrate to a domain: BiomedCLIP[16] pretrains on biomedical figure–caption pairs, Mammo-CLIP[17] on mammography, and BioCLIP[18] on organism imagery, the last pairing its release with an intrinsic analysis of *why* a domain substrate yields generalizing representations.

3. Overview

3.1. *Data*

Experimental studies are conducted on the VinDr-Mammo dataset,[10] which consists of 5,000 mammography studies. We apply a study-level 80:20 split with $n_{\text{train}} = 3{,}999$ and $n_{\text{test}} =$

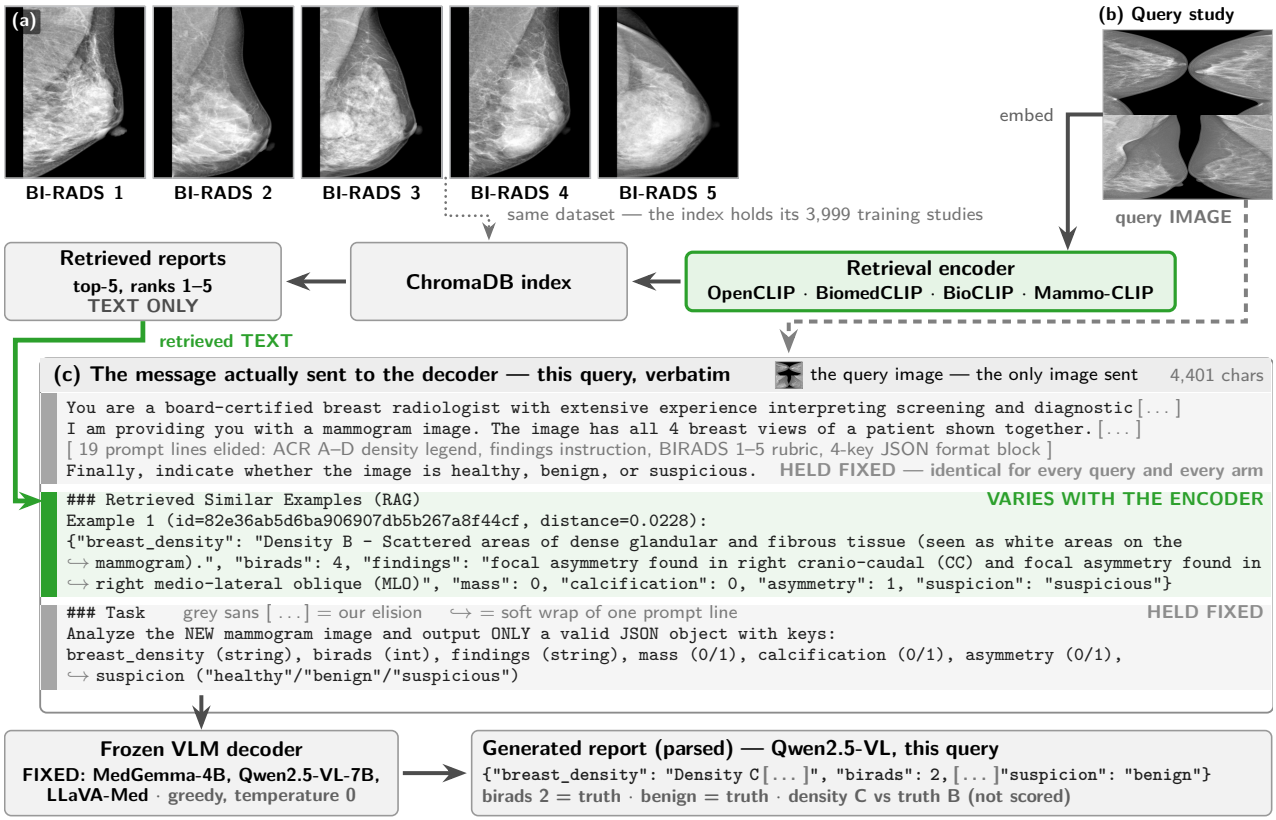


Fig. 1. The RAG Pipeline. **(a)** A representative of VinDr-Mammo studies, one per BI-RADS category, from whose density and finding annotations the structured ground-truth reports are synthesized. **(b)** The query study, a merged four-view composite and the image the decoder sees. **(c)** The message actually sent, verbatim for one query (Mammo-CLIP+SupCon → Qwen2.5-VL), elisions marked: gray is held fixed for every query and every arm, green is the retrieved block and the only part the encoder can change.

1,000 (one study is excluded for incomplete views: duplicate L-CC, missing L-MLO). VinDr-Mammo does not release patient identifiers, so the split cannot be made at the patient level. Ground-truth structured reports are synthesized from the dataset's annotations, with BI-RADS-derived suspicion mapping BI-RADS category 1 → healthy, 2–3 → benign, and 4–5 → suspicious. On the held-out test set, 487 of 1,000 studies are healthy, so a query-blind constant retriever attains P@1 = 0.487, the held-out prevalence of the majority class. We use this as the majority-class floor throughout the experiments. Suspicion is defined from BI-RADS categories, so the two are one clinical judgment at two granularities rather than independent tasks, and SupCon is trained on the suspicion labels. Reports contain seven generated fields. We analyze two endpoints, and the rest are generated but not scored. Predicted suspicion accuracy is always read from the model's own `suspicion` field, not from predicted BI-RADS categories. In the retrieval-layer suspicion, the top-5 plurality over the neighbors' own field is a separate encoder-level quantity and labeled as such throughout. The target is a radiologic assessment derived from BI-RADS, not a pathology endpoint, since VinDr-Mammo lacks biopsy confirmation for every case.

3.2. *Method*

The pipeline in Fig 1 shows the experimental setup. Each query mammogram is embedded by the retrieval encoder (e.g OpenCLIP, MammoCLIP, etc) into a vector, which then passes to ChromaDB to compute cosine similarity against the 3,999 training embeddings; the five highest cosine-score are retrieved.

After we construct the few-shot examples from the top-5 retrieved training studies (the five nearest the query image in embedding space), we feed the decoder their five structured JSON reports together with the query image and the fixed prompt template of Fig. 1(c). For each fixed decoder, we vary only the retrieval encoder, holding ChromaDB's retrieval mechanism, the prompt template, and the evaluator fixed.

For each encoder varied, we fine-tune each arm with the supervised-contrastive (SupCon) loss,[9] applied on the ℓ_2-normalized embeddings z_i using the three-class suspicion label y_i (healthy/benign/suspicious) and temperature τ:

$$L_{\text{SupCon}} = -\sum_i \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k \neq i} \exp(z_i \cdot z_k/\tau)}, \qquad P(i) = \{j : y_j = y_i,\ j \neq i\}. \tag{1}$$

The $1/|P(i)|$ normalization weights every positive equally, such that same-class batch members do not dominate.

3.3. *Encoder arms and substrates*

We evaluate four different encoders: OpenCLIP[5] (ViT-B/32, generic web images), BiomedCLIP[16] (ViT-B/16, biomedical figures), BioCLIP[18] (ViT-B/16, tree-of-life organism photos), and Mammo-CLIP[17] (EfficientNet-B5, mammography). Each is taken frozen and after the fine-tune of §3.4, giving eight arms, each crossed with the three frozen decoders. Each encoder runs at its own pretrained resolution: the three out-of-domain ViTs learn positional embeddings on a 224×224 patch grid[19] and ingest the merged four-view composite downscaled to 224×224, as in the pipeline we build on,[3] whereas Mammo-CLIP is fully convolutional, has no positional embeddings, and ingests 1520×912 single views. Their embeddings are correspondingly 512-d and 2048-d. The input resolution, view aggregation, and dimensionality therefore vary with architecture by construction rather than as factors set independently. BioCLIP shares OpenCLIP's and BiomedCLIP's architecture, dimensionality, input format, and fine-tune schedule, but differs in pretraining.

3.4. *Supervised-contrastive fine-tuning*

The SupCon objective, AdamW optimizer, cosine learning-rate schedule, and temperature $\tau = 0.07$ are identical across all encoder arms. For the three ViT encoders, we unfreeze the last $K = 6$ blocks ($\sim$43M trainable parameters) and train for 40 epochs at batch size 128 (learning rate 3×10^{-5}). Mammo-CLIP's higher-resolution EfficientNet-B5 tower is fine-tuned on the same 40-epoch, batch-128 schedule (last $K = 8$ blocks—matched to its greater depth—on 760×456 training views, half its native 1520×912 to fit the batch in memory). The objective, optimizer, learning-rate schedule, epoch count, and batch size are thus identical across all four arms. The unfrozen-block count and input resolution differ, due to the architecture. Both

states embed at native 1520×912, so the frozen-versus-fine-tuned contrast isolates fine-tuning from resolution. The training below inference resolution is a benign regime for convolutional networks.[20] Every arm selects the checkpoint with the best P@1 on a stratified 15% from the 3,999 training studies. SupCon fine-tuning therefore fits 3,401 studies, and the retrieval index at inference contains all 3,999. The 1,000 studies for the test set is evaluated once for the reported numbers.

3.5. *Retrieval geometry metrics*

On ℓ_2-normalized embeddings $z_i \in \mathbb{S}^{d-1}$ and the suspicion label y_i, we track two retrieval-geometry statistics. Precision P@1 is the fraction of query studies whose cosine-nearest neighbor shares the query's label: test-P@1 ranks each held-out query against its nearest *training* embedding, while train-P@1 ranks each training point against its nearest *other* training point. The angular class margin is the gap between mean same-class and mean different-class cosine similarity,

$$\Delta = s_{\text{intra}} - s_{\text{inter}}, \qquad s_{\text{intra}} = \mathbb{E}[\langle z_i, z_j \rangle \mid y_i = y_j], \quad s_{\text{inter}} = \mathbb{E}[\langle z_i, z_j \rangle \mid y_i \neq y_j], \tag{2}$$

and the ratio $\Delta_{\text{train}}/\Delta_{\text{test}}$ diagnoses overfitting. Δ is a supervised separation diagnostic. Within a fixed distribution, larger Δ monotonically increases P@k. The held-out caveat is that a train-set Δ increase need not transfer to held-out P@1 once the encoder is fit to the training set. We also report the total effective rank, the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the centered embedding covariance spectrum, which counts how many directions the representation actually uses. Note that P@1 and Δ are computed on the same three-class label the SupCon objective optimizes, so they measure held-out transfer of that objective, whereas the five-class BI-RADS endpoint is untrained.

3.6. *Report-level metrics and statistical tests*

After each of the three VLMs generates a report conditioned on retrieved examples from the varied encoders (frozen or fine-tuned), we evaluate accuracy performance. The baseline evaluated against is the majority-class predictor (what machine learning conventionally calls a dummy classifier) which scores 0.487 on the held-out split. At the retrieval layer of Fig. 1, where ChromaDB's index retrieves reports, we use precision at rank 1 (P@1) on the 1,000 held-out studies, counting a query correct when its top retrieved report shares its suspicion label. For each encoder in both states, frozen and fine-tuned, we test held-out P@1 against the dummy classifier with a one-sided binomial test and report a 95% Wilson interval alongside, so that we can tell whether an arm lands above the baseline, level with it, or significantly below. These tests would allow us to quantify how likely an observed difference is to have increased from sampling variation alone, answering whether a specific encoder is better than a dummy classifier. Furthermore, we apply exact paired McNemar tests on the shared 1,000 to queries compare each fine-tuned encoder with its frozen state. We also computed the pair-bootstrap on the raw experimental results to statistically test the change. The 95% percentile interval of the difference is reported. The frozen-versus-fine-tuned design yields 28 such contrasts (four

Table 1. The retrieval-to-report results for different Encoder–State combinations.

Encoder	State	Train	Transfer to held-out retrieval					Report Δacc (rel. %)		
		train P@1	Δ_{tr}/ Δ_{te}	eff. rank	test P@1	vs. floor	susp. vote	Qwen 2.5-VL	LLaVA -Med	MedGemma
OpenCLIP	frozen	0.455	1.14	10.53	0.432	below	0.136	—	—	—
	+SupCon	**0.510**	4.73	1.10	**0.458**	tied	0.214	−2.6	−2.2	+2.0
BiomedCLIP	frozen	0.450	0.94	9.07	0.442	below	0.107	—	—	—
	+SupCon	**0.547**	4.81	2.37	**0.491***	tied	0.204	+11.0*	+7.9	+16.4†
BioCLIP	frozen	0.458	1.13	20.15	0.419	below	0.136	—	—	—
	+SupCon	**0.546**	5.11	2.47	**0.464***	tied	0.252	+3.7	+6.6	+2.0
Mammo-CLIP	frozen	0.519	1.26	7.20	0.513	tied	0.146	—	—	—
	+SupCon	**0.578**	1.15	13.25	**0.585**†	above	**0.291**	+**20.7**†	+**15.6**†	+4.0

Δ is the angular margin of Eq. (2) and is *not* comparable across the 512-d ViT and 2048-d convolutional families, whereas the ratio and P@1 are comparable across both. *eff. rank* is the participation ratio on the held-out split. *vs. floor* is the verdict of the 95% Wilson interval on test-P@1 against 0.487—*below*, *tied* or *above* as it falls under, within, or above the floor. *susp. vote* is the five retrieved neighbors' majority label scored on the 103 suspicious queries. As a per-class rate, it is not comparable with the floor. Report Δacc is (SupCon − frozen)/frozen × 100 on suspicion accuracy (Figs. 2–3), not comparable across cells or decoders (LLaVA-Med served 4-bit against bfloat16, constant within each decoder). *One marker convention throughout*: $^{*}p < 0.05$, $^{\dagger}p < 0.001$, exact paired McNemar frozen vs. +SupCon, uncorrected.

at the retrieval layer plus three decoders × four encoders × two tasks — suspicion and BI-RADs), so we control multiplicity across this pre-specified family with Holm–Bonferroni and Benjamini–Hochberg.

4. Results

4.1. *Separation without transfer: only the domain-native encoder generalizes*

Table 1 evaluates the held-out nearest-neighbor retrieval against the majority-class floor of P@1 = 0.487. The held-out quantities are on the same n = 1,000 studies, and train-P@1 are on the 3,999 training studies. The main finding is not that supervised contrastive learning uniformly improves mammography RAG, but that its transfer depends strongly on the representation. The frozen-to-SupCon comparisons isolate changes induced by the retrieval encoder. In the frozen states, OpenCLIP, BiomedCLIP, and BioCLIP obtain test P@1 values of 0.432, 0.442, and 0.419, respectively. For each encoder, the 95% Wilson interval lies entirely below the floor. Under the common SupCon objective and matched optimizer, learning-rate schedule, epoch count, and batch size, the held-out P@1 increases to 0.458, 0.491, and 0.464 for OpenCLIP, BiomedCLIP, and BioCLIP, respectively. BiomedCLIP and BioCLIP improve

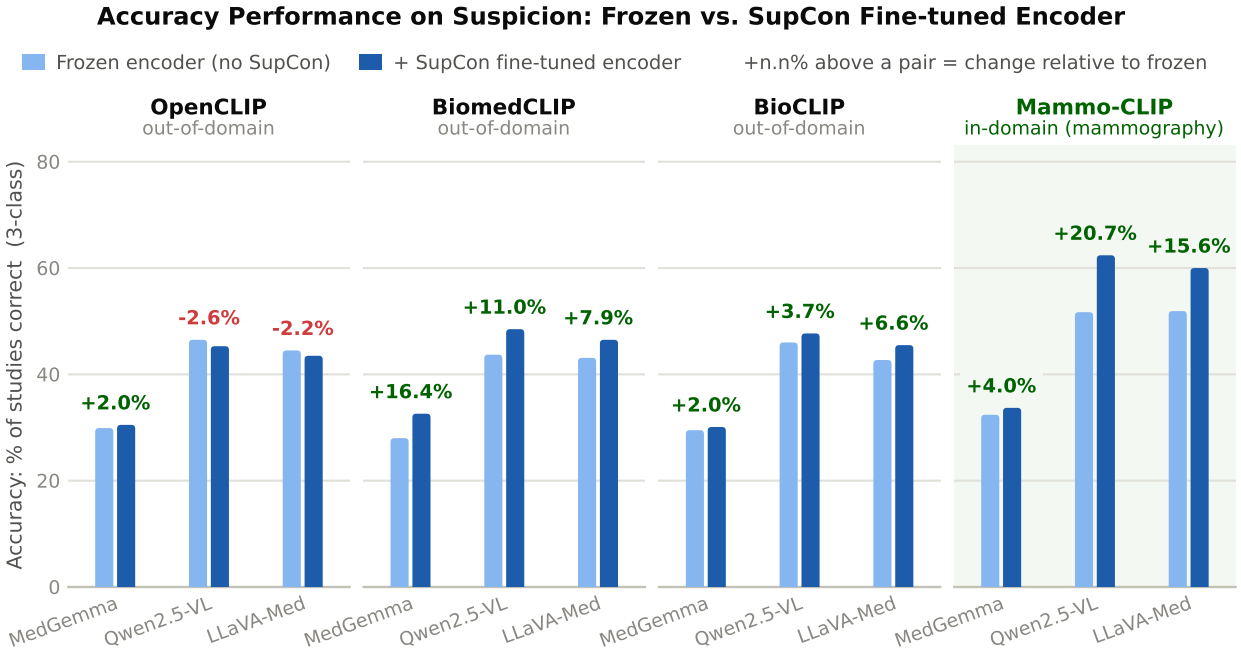


Fig. 2. Evaluation on the suspicion label (3-class).

significantly relative to their own frozen states under the unadjusted paired McNemar tests, showing that fine-tuning does improve retrieval outcomes. However, the 95% Wilson interval for every tuned out-of-domain encoder still overlaps the 0.487 floor. Thus, fine-tuning improves some out-of-domain encoders relative to baseline but does not establish above-floor held-out retrieval. Fine-tuning does exactly what the objective asks on the training set, inflating the train-pair angular margin of Eq. (2) by 9–23× across the three out-of-domain arms, while the held-out margin inflates far less, by only 2–5.5×; the resulting tuned-state ratio is tabulated directly (Table 1). The tuned $\Delta_{\text{train}}/\Delta_{\text{test}}$ ratio consequently increases from 0.94–1.14 to 4.73–5.11, while held-out effective rank decreases from 9.07–20.15 to 1.10–2.47. This pattern is consistent with a low-dimensional, training-specific solution rather than a transferable retrieval structure. For Mammo-CLIP, the test P@1 increases from 0.513 to 0.585, with the tuned encoder significantly above the majority-class floor. The train/test margin ratio decreases from 1.26 to 1.15, and the effective rank increases from 7.20 to 13.25. These results demonstrate the training-set separation also transfers to held-out retrieval.

Every quantity above is measured upstream of the decoder, at the retrieved-reports stage of Fig. 1. Because the objective, optimizer, schedule, epoch count, and batch size are identical across all four arms, these shared settings do not explain the difference in outcome between Mammo-CLIP and the other three encoders. Separation alone is therefore not what matters: whether it generalizes depends on the encoder. This experiment shows that the choice of encoder decides whether separation generalizes at the retrieval layer. We next ask whether it has any downstream impact.

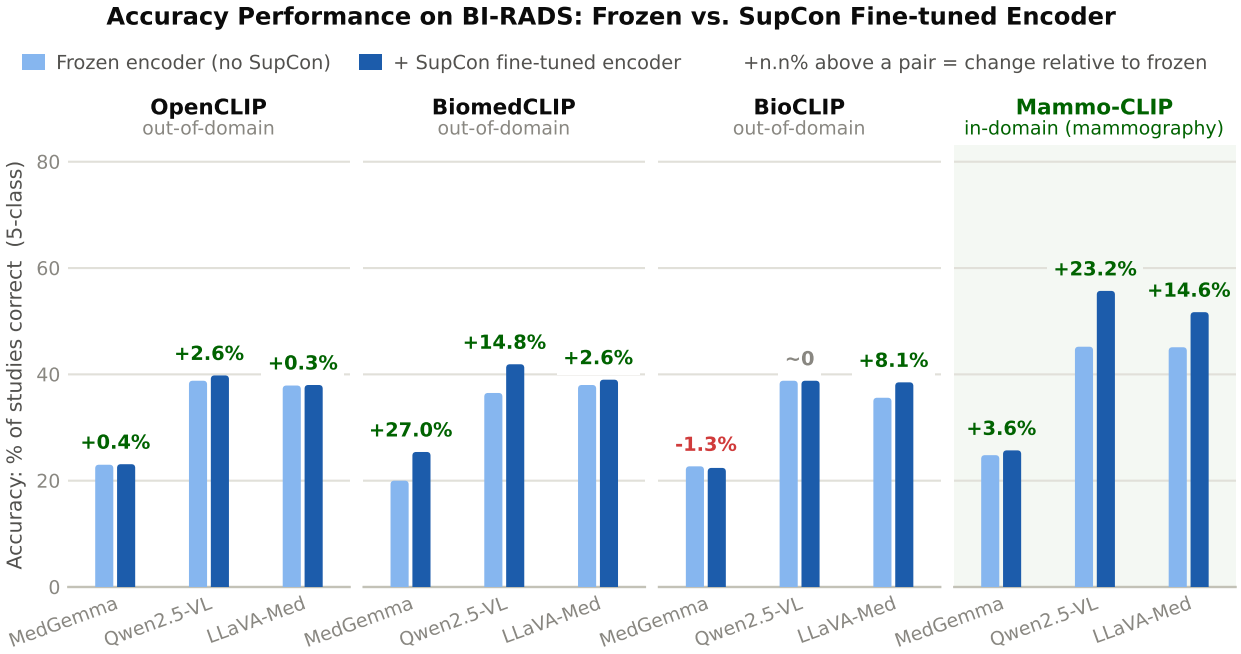


Fig. 3. Evaluation on the BI-RADS label (5-class).

4.2. *Improvement without clinical reach: the gain concentrates in the routine band*

Figs. 2 and 3 indicate that on both the suspicion and BI-RADS tasks, MedGemma paired with BiomedCLIP shows a large gain from frozen to fine-tuned (+16.4% and +27.0%). In terms of magnitude, MedGemma's frozen baselines are the lowest of the three decoders (0.280 on suspicion, 0.200 on BI-RADS), so those relative gains are only +0.046 and +0.054 in absolute terms, small but significant. Report-level gains therefore do not require above-floor retrieval. For MedGemma, the gain depends on the specific encoder, not on the encoder's retrieval quality: Mammo-CLIP, the only encoder above the floor, changes MedGemma's accuracy by only +4.0% on suspicion and +3.6% on BI-RADS, both not significant. Note that this results does not directly explain why MedGemma responds to BiomedCLIP. It is possible that P@1 checks only whether a retrieved report carries the same suspicion label as the query, whereas the decoder reads the whole report (breast density, findings, mass, calcification). A report with the wrong suspicion label can still describe a breast that resembles the query, and that text may help MedGemma even though P@1 scores it a miss.

To make the evaluation clinically relevant, we introduce ACR management band,[21] in which we group the true BI-RADs categories by the action a radiologist would take: routine screening (BI-RADS 1–2), short-interval follow-up (BI-RADS 3), and biopsy (BI-RADS 4–5). Fig. 4 decomposes the largest gain in the grid, Mammo-CLIP into Qwen2.5-VL. Of the +107 studies gained overall, +91 fall in the routine band (0.595 → 0.708, $n = 807$), +10 in short-interval follow-up (0.400 → 0.511, $n = 90$, not significant), and +6 in the biopsy band (0.010 → 0.068, $n = 103$, not significant). Routine and follow-up studies improved at nearly

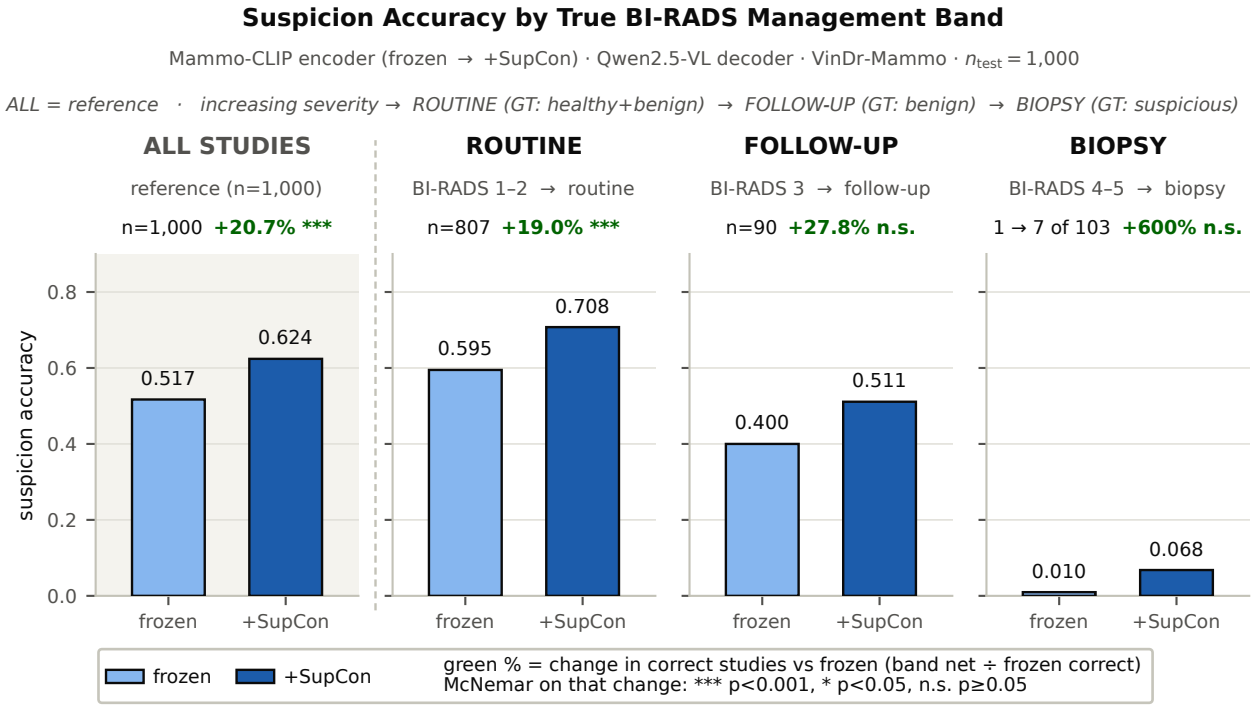


Fig. 4. The suspicion accuracy performance categorized into routine screening $\{1, 2\}$, short-interval follow-up $\{3\}$ and biopsy $\{4, 5\}$. The biopsy bands are post-hoc the p-values in green are unadjusted. VinDr-Mammo labels are radiologist assessments, so the biopsy band denotes a recommendation for tissue sampling, not confirmed malignancy.

the same rate (+0.113 and +0.111), but only the routine band reaches significance. In the biopsy band, the fine-tuned model identifies seven out of 103 studies against one when frozen. The report level thus repeats the pattern of the retrieval level: the quantity being optimized improves overall, and the quantity that matters clinically does not. That is why held-out precision against the prevalence floor is reported rather than the training-set margin, and the management-band decomposition rather than aggregate accuracy alone.

5. Limitations

The VinDr-Mammo dataset analyzed in the experiments is highly imbalanced with a small sample size in the higher BI-RADS categories (4 and 5). Thus, from the results, the accuracy performance improvements are clearly seen and concentrated on more healthy and benign studies. Based on the suspicion class, the number is small on Qwen2.5-VL (7 studies newly correct against 1 newly wrong, net +6 of 103) and indistinguishable from none on LLaVA-Med (9 against 8, net +1). Additional mammography datasets, class-balanced sampling, and aggregation across pretraining domains would further assess robustness. LLaVA-Med is served four-bit-quantized, whereas MedGemma and Qwen2.5-VL use bfloat16. Serving is fixed within each decoder, preserving the paired encoder contrasts, but cross-decoder magnitudes are not directly comparable. Additionally, the SupCon positive set is defined on suspicion labels, which treats BI-RADS 2 as equivalent to 3 and 4 as equivalent to 5 and supplies no gradient on the

within-block distinction, which ordinal contrastive learning could address.

Additionally, Mammo-CLIP may differ from the three out-of-domain encoders beyond its pretraining corpus, including architecture (EfficientNet-B5 versus ViT-B), embedding dimensionality (2048 versus 512), input resolution and view handling (1520×912 single views versus a 224×224 four-view composite), and unfrozen blocks ($K = 8$ versus $K = 6$). These differences could limit the ability to attribute the observed transfer advantage specifically to domain-specific pretraining. Future work could assess these factors through more closely controlled comparisons and clarify whether Mammo-CLIP's transfer advantage is primarily attributable to domain-specific pretraining or to architectural and input-processing differences, and whether general-purpose pretraining followed by in-domain fine-tuning can provide a viable alternative to a domain-native encoder.

6. Conclusion

This work focuses on a modality-specific bottleneck in retrieval-augmented mammography reporting: the visual retrieval step, which prior work held fixed while tuning the language decoder. Fine-tuning a varied encoder for every fixed frozen decoder helps to assess whether separating class geometry in embedding space improves the accuracy of generated reports in downstream VLMs. When retrieval is image-conditioned, the choice of pretrained encoder determines whether retrieval generalizes. Because pretraining corpus, architecture, and input vary in the design, whether the pretraining substrate is the cause remains open for further research. The supervised-contrastive fine-tuning drives the three out-of-domain encoders toward a low-rank geometry that separates the training set while leaving held-out retrieval tied with the majority-class floor, and expands the mammography-native one into a representation that results in performance gains on the generated report. With the decoder held fixed, substantial gains are therefore available from the retrieval space alone. The result provides grounded insight and guides future work on clinically meaningful contrastive pairs and retrieval-database curation.

Code Availability

Code for encoder fine-tuning, retrieval, report generation, and the auditing protocol will be released at `https://github.com/dcnguyen060899/separation-without-transfer` upon publication. The VinDr-Mammo Dataset on PhysioNet is governed by the PhysioNet Restricted Health Data, thus the code released is rebuilt from a licensed copy, not the retrieval index.

References

1. X. Zhang, J. N. Acosta, X. Yang, S. Adithan, L. Luo, H.-Y. Zhou, J. Miller, O. Huang, Z. Zhou, I. E. Hamamci, S. Bannur, K. Bouzid, X. Zhang, Z. Meng, A. Nicolson, B. Koopman, I. Baek, H. Ko, M. P. Ranjit, S. Srivastav, S. G. Sambanthan and P. Rajpurkar, Automated chest x-ray report generation remains unsolved, in *Pacific Symposium on Biocomputing*, (World Scientific, 2026).
2. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t.

Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems (NeurIPS)* **33** (2020).
3. R. Jahangir, N. I. Khan, A. Sudheerkumar and V. Filkov, Mammowise: A multi-model local rag pipeline for mammography report generation, *arXiv preprint arXiv:2602.22462* (2026), arXiv:2602.22462v1.
4. T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
5. M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt and J. Jitsev, Reproducible scaling laws for contrastive language-image learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
6. Google DeepMind, Medgemma: A family of medical vision-language models, *Technical Report* (2024).
7. Qwen Team, Qwen2.5-vl technical report, *arXiv preprint arXiv:2409.12191* (2024).
8. C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon and J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
9. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, Supervised contrastive learning, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
10. H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao and V. Vu, Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography, in *Scientific Data*, (1) (Nature Publishing Group, 2023).
11. A. Lozano, S. L. Fleming, C.-C. Chiang and N. Shah, Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature, in *Pacific Symposium on Biocomputing*, (World Scientific, 2024).
12. G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu and A. Zhang, Improving retrieval-augmented generation in medicine with iterative follow-up questions, in *Pacific Symposium on Biocomputing*, (World Scientific, 2025).
13. G. Xiong, Q. Jin, Z. Lu and A. Zhang, Benchmarking retrieval-augmented generation for medicine, in *Findings of the Association for Computational Linguistics: ACL 2024* , (Association for Computational Linguistics, 2024).
14. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning transferable visual models from natural language supervision, in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
15. T. Wang and P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
16. S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang and H. Poon, Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, *arXiv preprint arXiv:2303.00915* (2023), Also NEJM AI, 2025.
17. S. Ghosh, C. B. Poynton, S. Visweswaran and K. Batmanghelich, Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* , (Springer, 2024). arXiv:2405.12255.
18. S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong,

W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao and Y. Su, Bioclip: A vision foundation model for the tree of life, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
19. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations (ICLR)*, 2021.
20. H. Touvron, A. Vedaldi, M. Douze and H. Jégou, Fixing the train-test resolution discrepancy, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
21. E. A. Sickles, C. J. D'Orsi, L. W. Bassett *et al.*, *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, 5 edn. (American College of Radiology, Reston, VA, 2013).