ZERO-SHOT CONTEXT-AWARE MEDICAL CASE RETRIEVAL VIA TOKEN-SET TO TOKEN-SET SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In healthcare, model-predicted outcome labels are, in practice, unhelpful, especially when clinicians are concerned about the uncertainty of these predictions. In contrast, references to past similar cases provide valuable guidance. Therefore, in this work, when presented with a new patient, our objective is to retrieve similar past medical cases to support treatment planning within a specific clinical context of interest. Previous methods rely on semantic searches and content-based image retrieval using vector databases of EHR and medical image embeddings. Although efficient, these approaches often lack contextual relevance, leading to lower retrieval performance. Recent multimodal foundation models can enable more effective context-aware cross-modal retrieval. While these models provide deeper context and richer embeddings, they are computationally expensive. Considering this, we propose a two-stage retrieval system that combines the efficiency of vector-based searches with the contextual depth of token embeddings provided by foundation models without further training. More importantly, we propose a token-set to token-set similarity for context-aware retrieval using rich token embeddings. Our extended experiments demonstrate the efficiency contributed by Stage 1 and the effectiveness contributed by the context-aware re-ranking in Stage 2.

1 INTRODUCTION

Taking Chest X-rays (CXRs) as an example in healthcare, they are critical initial diagnostic tests in clinical practice, widely used due to their cost-effectiveness and accessibility, with over 1.4 billion performed annually (Chen et al., 2024b). Let's consider a common clinical scenario where a new patient is encountered with a chest X-ray and an associated radiology report. Under a specific diagnostic context of interest, the clinical team must make treatment decisions. While existing AI models can provide binary predictions for these conditions, these labels offer limited utility when clinical teams have significant concerns about prediction uncertainty. In contrast, similar past medical cases under the same context from a historical database provides invaluable references for confident treatment planning, allowing clinicians to examine how analogous presentations were diagnosed and managed (Akash R. J. et al., 2024), which is especially helpful for ambiguous cases.

In existing systems, practitioners often use Picture Archiving and Communication Systems (PACS) to retrieve and compare prior imaging studies (Intelerad, 2022). However, current systems lack sophisticated semantic search capabilities, making it resource-intensive to find truly analogous cases. Formally, given a query and diagnostic context, our goal is to retrieve the most relevant historical cases that share similar pathological characteristics under a specific context, serving as references for clinical decision-making. Content-Based Medical Image Retrieval (CBMIR) systems have evolved from hand-crafted features (Kumar et al., 2013) to deep learning approaches using vision transformers (Dosovitskiy et al., 2021). Many of these approaches require supervised training on labeled datasets, limiting their adaptability to new diagnostic contexts without retraining.

Vector database systems offer computational efficiency through approximate nearest neighbor search, enabling real-time retrieval from large medical archives. Contrastive learning models like CLIP (Radford et al., 2021) and medical variants produce aligned text and image embeddings, enabling cross-modal retrieval where clinical text can retrieve relevant images and vice versa. While efficient, these embedding-based approaches often lack the fine-grained discriminative power needed for nuanced medical distinctions, particularly when pathological findings are subtle or localized. Retrieval-

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.

Augmented Generation (RAG) has emerged as a promising paradigm for grounding large language model outputs in retrieved knowledge (Lewis et al., 2021; Chu et al., 2025; Xia et al., 2025). These methods demonstrate improved factual accuracy by incorporating retrieved medical images and reports. However, RAG fundamentally addresses a different problem: enhancing text generation quality rather than retrieving for clinical reference.

Recent multimodal medical foundation models (Chen et al., 2024b; Sellergren et al., 2025; Chen et al., 2024b; Xu et al., 2023) have demonstrated remarkable capabilities in medical reasoning, excelling at report generation, question answering, and disease classification. These models generate rich token-level representations that capture fine-grained relationships between visual findings and clinical concepts. Critically, these models encode domain knowledge that enables zero-shot adaptation to new diagnostic contexts through natural language specification. This allows the context c to be expressed as a textual query without model fine-tuning. This capability motivates a key insight: multimodal medical foundation models can enable context-aware medical case retrieval, where the diagnostic context guides the retrieval process toward cases most relevant to the specific clinical question at hand. Moreover, the rich knowledge embedded in these pretrained models offers the potential for zero-shot approaches, which is well desired in the healthcare practice with huge varieties of context.

Motivated by these considerations, we propose a two-stage zero-shot context-aware medical case retrieval framework. Stage 1 performs vector similarity search to rapidly generate a small pool of candidates from the large historical database for efficiency, while Stage 2 re-ranks candidates using a context-aware token-set to token-set similarity based on rich token embeddings provided by a multimodal medical foundation model without further training. Our main contributions are

1) We propose a novel zero-shot retrieval framework that takes advantage of the computational efficiency in vector database search and the semantic precision in token-set-based re-ranking;

2) We introduce a context-aware token-set to token-set similarity for re-ranking using rich token embeddings provided by multimodal medical foundation models without further training;

3) We conduct comprehensive experiments across six pathologies, demonstrating the efficiency and effectiveness of the proposed framework.

2 Related Work

In this section, we briefly review the existing literature related to similar medical case retrieval involving patient image and text data.

2.1 Content-Based Image Retrieval (CBIR)

Traditional CBIR systems in radiology focus on pixel-level similarity, which often does not reflect clinical relevance or disease severity (Kumar et al., 2013). Modern CBIR architectures approach this by learning pathological features. For example, Multi-Level Dual-Attention Mechanisms (ML-DAM) (Fang et al., 2021) combine Position Attention Modules (PAM) and Channel Attention Modules (CAM) to strengthen features at spatial and channel levels for a targeted pathology. (Silva et al., 2022) trains the network with saliency maps as direct inputs, which maximizes attention to disease-related regions. CheXtriev (Akash R. J. et al., 2024) introduces graph transformers to encode spatial context and anatomical relationships. Although these supervised CBIR methods fit well with clinical needs, they require pathology-specific training on labeled datasets and thus each pathology requires a separate feature extractor that may not generalize well. Moreover, they cannot take available textual information about the patient into consideration. In contrast, our approach operates in a zero-shot setting without further training and can consider both image and text information.

2.2 Cross-Modal Medical Retrieval

To involve textual information, medical datasets such as MIMIC-CXR (Johnson et al., 2019a; 2024), CheXpert Plus (Chambon et al., 2024), Chest ImaGenome (Wu et al., 2021), and CheXBench (Chen et al., 2024b) have enabled the development of medical adaptations of CLIP by compiled image-text pairs and even other forms of context like bounding boxes. Recent architectures such as MedCLIP (Wang et al., 2022), BiomedCLIP (Zhang et al., 2024), MedSigLIP (Sellergren et al.,

2025), XraySigLIP (Chen et al., 2024a), and CheXficient (Wang et al., 2026), use large-scale vision-language encoders to align visual tokens with clinical terminology, which can support cross-modal image-text retrieval. Recently, the CRMR framework (Fang et al., 2025) uses loss functions like HIP-InfoNCE (Dong et al., 2025) to align holistic images with stochastic masked views of reports, improving precision in dense clinical latent spaces. M^3Bind (Liu et al., 2025), on the other hand, extends multimodal alignment to five medical imaging modalities (X-ray, ECG, retina, pathology, and CT) using adaptive modality balancing and shared embedding space knowledge distillation. A key limitation of these embedding-based approaches is their use in global similarity scores, which can lack the fine-grained discrimination needed for subtle or localized medical findings.

Recently, RadIR (Zhang et al., 2025) proposed refining by applying RaTEScore (Zhao et al., 2024) to measure text-level similarity as a proxy for image similarity, in which a triplet loss can be used to learn a fusion module for anatomy-conditioned retrieval via text. Although RadIR achieves better retrieval performance, it requires additional supervised training on anatomy-annotated data that is not often accessible and is non-trivial to create without the help of board-certified physicians.

2.3 Multimodal Medical LLMs & RAG

CheXagent (Chen et al., 2024b) and other multimodal medical LLMs (Sellergren et al., 2025; Xu et al., 2023; Bannur et al., 2024) excel in medical reasoning, report generation, visual question answering, and diagnostic classification by producing rich token-level representations that grasp fine-grained relationships between visual findings and clinical concepts. These models encode domain knowledge that enables zero-shot capabilities in new diagnostic contexts through natural language prompts, allowing clinical queries as free-text without model fine-tuning. While multimodal medical LLMs perform well in reasoning and interpretation, they continue to be prone to hallucinations. To address this, RAG-based (Lewis et al., 2021) systems supplement parametric knowledge by retrieving similar cases during inference. Recently, Visual-RAG (V-RAG) (Chu et al., 2025), MMed-RAG (Xia et al., 2025), and LaB-RAG (Song et al., 2025) have demonstrated that incorporating retrieved text and visual data from similar images improves entity probing accuracy and reduces hallucinations in chest X-ray report generation. Note that during retrieval that is related to our work, almost all of them rely on pre-trained or fine-tuned embeddings, as discussed in the above subsection. Although multimodal medical LLMs are not designed for retrieval tasks, they generate informative token-level representations that can capture fine-grained context for context-aware retrieval without further training, which is explored in this work and can also be used in RAG.

2.4 Token-Level Late Interaction and Optimal Transport

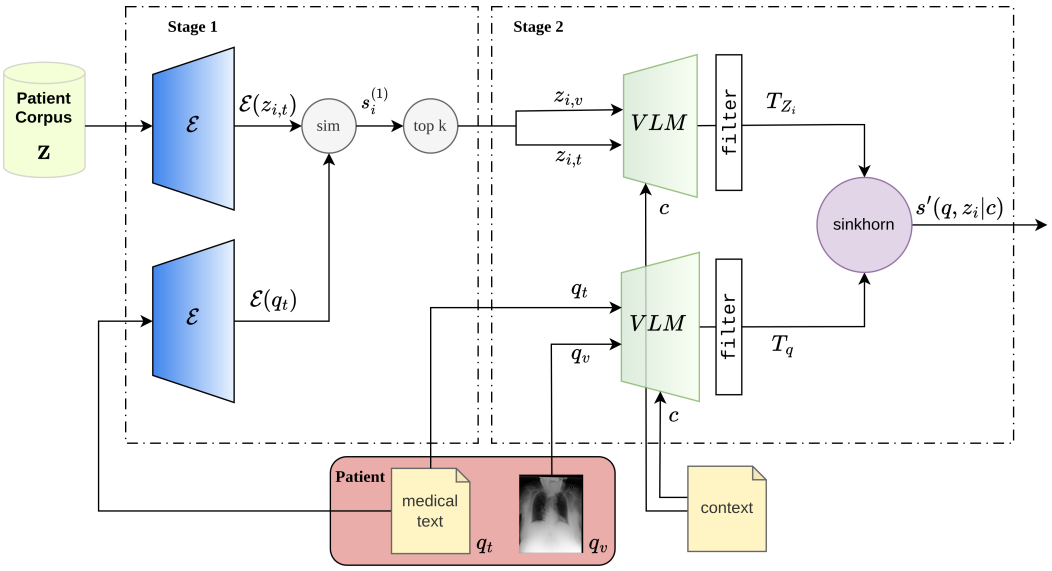
Beyond global embedding similarity, a growing body of work computes fine-grained similarity over individual token representations. ColBERT Khattab & Zaharia (2020) introduced the late interaction paradigm, in which queries and documents are independently encoded into per-token embeddings, and relevance is scored via a MaxSim operator that matches each query token to its most similar document token. Recent work extends late interaction to vision-language and medical settings. ColPali Faysse et al. (2025) applies MaxSim over PaliGemma patch embeddings for OCR-free visual document retrieval, demonstrating that VLM-derived multi-vector representations can replace traditional text-based pipelines. Several medical adaptations have followed, including C-MIR Jush et al. (2025) for 3D volumetric image retrieval and RadLIT Matulich (2026) for radiology education, both building on ColBERT-style architectures with domain-specific encoders.

A related line of work replaces the MaxSim operator with optimal transport (OT) for structured set-to-set alignment, capturing many-to-many correspondences that MaxSim's greedy one-to-one matching cannot represent. For example, PLOT Chen et al. (2023) uses that to align multiple text prompts against local visual features during model training, while MOTOR Shaaban et al. (2025) adopts that to improve medical visual question answering. In this work, we focus on set-to-set similarity for retrieval, which is especially context-aware.

3 The Proposed Method

Given a paired input of $q = \{q_v, q_t\}$ for a new patient's image and text data, respectively, our objective is to retrieve some most similar medical records in corpus $\mathbf{Z}$ under a specific medical

Figure 1: Our retrieval first performs a vector-based search given embeddings of the new patient and the corpus with encoder $\mathcal{E}$ (text embeddings as an example) to generate k candidates. Then, an existing VLM generates informative tokens with only context-aware ones go through the filter to calculate token-set to token-set similarity, $sinkhorn$, for fine-grained re-ranking.



concern, denoted as context c. The context c is provided by the clinician at inference time as a free-text diagnostic query. $\mathbf{Z}$ consists of medical records $\{z_{i,v}, z_{i,t}\}\ \forall z_i \in \mathbf{Z}$ corresponding to the image and text data of a diagnosed patient, respectively. The main challenge is to balance efficiency and effectiveness. Considering that a sophisticated retrieval algorithm involving both image and text data can be very expensive when applied to a large corpus, we propose a two-stage method as follows.

3.1 Stage 1: Efficient Coarse-grained Retrieval with Fixed Embeddings

Given a new patient record q and previous records in $\mathbf{Z}$, we first perform an efficient coarse-grained similarity search. We adopt the cosine similarity defined on fixed text or image embeddings to efficiently obtain candidate medical records from the large corpus as the first stage.

Specifically, taking the text embeddings as an example, let $\mathcal{E}$ denote a pretrained text encoder. The representations for the query patient q_t and the record $z_{i,t}$ can be extracted as $\mathcal{E}(q_t)$ and $\mathcal{E}(z_{i,t})$, respectively. Then, the cosine similarity between them can be obtained as

$$s_i^{(1)} = \frac{\mathcal{E}(q_t)^\top \mathcal{E}(z_{i,t})}{\|\mathcal{E}(q_t)\| \|\mathcal{E}(z_{i,t})\|} \tag{1}$$

Image embeddings can be used similarly. It should be noted that this embedding-based similarity search can be implemented efficiently using FAISS (Douze et al., 2024; Johnson et al., 2019b). While fixed text or image embeddings can summarize the key observations for each patient, the fine-grained details relevant to context are hard to retain. To improve the retrieval performance, we select a small number of candidates and then pass them to the next stage for refinement with fine-grained details.

3.2 Stage 2: Fine-grained Re-ranking by Token-set to Token-set Similarity

We then refine the ranking of the obtained candidates, $\mathbf{Z}^{(1)}$, by exploiting additional details conditioned on the target medical context c. To enable that, we first investigate the architecture of popular pretrained models. We find that both the vision and text encoders can apply transformers as the backbone to extract representations. While the feature of the `[CLS]` token is adopted as the

representation for the embedding sequence, transformers also generate tokens for image patches and text words, which retain sufficient details about the original image or text.

To exploit the detailed information for the second stage, we leverage the additional tokens to compute the similarity. The main challenges are two aspects: modality alignment for multimodal tokens and similarity measurement defined on token-set to token-set.

3.2.1 MODALITY-ALIGNED TOKENS

Since the tokens can be from images or text reports, the representation between vision and text tokens should be aligned to compute the similarity over multiple modalities. While CLIP-style models align vision and text tokens by contrastive pre-training, the learning objective is defined only on `[CLS]` tokens, where the optimization for detailed tokens can be insufficient.

In contrast, decoder-style multimodal LLMs rely on vision and text tokens to handle diverse tasks, including visual question answering (VQA), which requires the appropriate details to respond correctly. Because both image patch tokens and text tokens are concatenated into a single sequence processed by a causal transformer, the final-layer hidden states at text token positions have already attended to all preceding image and text tokens through self-attention. We leverage this property to extract token-level representations that capture fine-grained context without further training.

Stage 2 requires an existing model that satisfies three properties: 1) it accepts a joint multimodal input consisting of a image, a text report, and a context prompt in a single forward pass; 2) it produces per-token hidden states at a specified decoder layer, where text token representations have attended to preceding image and text tokens through the model's self-attention mechanism; and 3) these hidden states are extractable at inference time without generation. The model is used as a feature extractor by running a forward pass and reading off internal representations rather than auto-regressively decoding output tokens. These requirements are satisfied by the broad class of decoder-style medical VLMs that process visual and textual tokens jointly.

Let $\text{VLM}(\cdot)$ denote a pretrained model satisfying the above properties. The VLM takes an image, its corresponding text report, and a medical context question as inputs and produces token representations through its vision and language encoders, which feed into a transformer decoder. We rewrite the medical context c as "Does this patient have c?" before feeding it to the VLM, for the sake of simplicity in this work.

For each record $z_i \in \mathbf{Z}$, we feed the image $z_{i,v}$, text report $z_{i,t}$, and medical context question c jointly through the VLM and take the hidden states from the final layer of the transformer decoder. Since these output tokens have attended to the image and text report through the decoder's self-attention, the token positions corresponding to c encode a context-aware representation of the record with respect to the medical question of interest. We extract the subset of output tokens at the positions aligned with c:

$$T_{z_i} = \text{VLM}(z_{i,v}, z_{i,t}, c)\big|_c \in \mathbb{R}^{L_c \times d}$$

where L_c is the number of tokens in the medical context question and d denotes the feature dimension. Each token is normalized to have unit norm. Similarly, for the query patient, we obtain

$$T_q = \text{VLM}(q_v, q_t, c)\big|_c \in \mathbb{R}^{L_c \times d}$$

which represents the query's context-aware tokens with respect to the same medical question.

3.2.2 SIMILARITY MEASUREMENT FROM TOKEN-SET TO TOKEN-SET

Compared with the single feature vector for the query and candidates in the first stage, we observe a set of tokens for each record to retain the fine-grained details in this second stage. Measuring the similarity between two token sets requires a matching function that can compare all pairs of tokens and aggregate them into a scalar score. Inspired by the theoretical work on optimal transport (OT) (Rubner et al., 2000; Chuang et al., 2023; Chen et al., 2020; Huang et al., 2023) and its recent applications in model training and QA (Khattab & Zaharia, 2020; Faysse et al., 2025), we introduce a token-set to token-set similarity as follows for retrieval, which is not a greedy one-to-one matching like MaxSim but a distribution-to-distribution matching without any token bias.

$$s(q, z_i|c) = \max_{P \in \Delta} \langle P, M \rangle \tag{2}$$

where $\Delta = \{P \in \mathbb{R}_+^{L_q \times L_{z_i}} | P\mathbf{1}_{L_q} = 1/L_{z_i}, P^\top \mathbf{1}_{L_{z_i}} = 1/L_p\}$ and $M = T_q T_{z_i}^\top$. As noted above, in all configurations evaluated in this work $L_q = L_{z_i}$, so M is square; the general formulation is retained for extensibility.

The similarity in Eqn. 2 can be obtained by solving the linear programming problem. To further reduce the computational complexity, an additional constraint can be introduced as in the Sinkhorn distance (Cuturi, 2013).

$$s'(q, z_i|c) = \max_{P \in \Delta_\alpha} \langle P, M \rangle \tag{3}$$

where $\Delta_\alpha = \{P \in \Delta | H(P) \geq \alpha\}$ and $H(\cdot)$ computes the entropy in P. Compared with Eqn. 2, Eqn. 3 introduces the entropy constraint to makes the transport matrix P smooth that can be more robust to the input similarity M. In addition, according to the dual theory (Boyd & Vandenberghe, 2014), the problem in Eqn. 3 is equivalent to

$$\max_{P \in \Delta} \langle P, M \rangle + \tau H(P) \tag{4}$$

which can be solved efficiently by Sinkhorn's fixed-point iteration (Cuturi, 2013). The detailed algorithm is summarized in Alg. 1.

Algorithm 1 Token-set to Token-set Similarity via Sinkhorn

Require: Similarity matrix $M \in \mathbb{R}^{m \times n}$, temperature τ, iterations T
1: $P \leftarrow \text{softmax}(M/\tau, \text{dim} = 1)$ {Row-wise softmax}
2: $P \leftarrow P/m$
3: **for** $t = 1$ to T **do**
4: $\quad c_j \leftarrow \sum_{i=1}^{m} P_{ij} \quad \forall j$ {Column normalization}
5: $\quad P_{ij} \leftarrow \frac{P_{ij}}{c_j \cdot n} \quad \forall i, j$
6: $\quad r_i \leftarrow \sum_{j=1}^{n} P_{ij} \quad \forall i$ {Row normalization}
7: $\quad P_{ij} \leftarrow \frac{P_{ij}}{r_i \cdot m} \quad \forall i, j$
8: **end for**
9: **Returns** $\sum_{i,j} P_{ij} \cdot M_{ij}$

After obtaining the transport matrix P by the algorithm, the similarity between query and candidate records can be computed as in Eqn. 3. The proposed retrieval framework is illustrated in Fig. 1, in which text embeddings are used in Stage 1 as an example.

4 Experiments

We evaluate our two-stage retrieval framework through a series of experiments designed to answer three main questions, while other details can be found in the appendix: 1) Which pretrained encoder provides the strongest Stage 1 baseline? 2) Does Stage 2 re-ranking with token-set to token-set similarity improve retrieval quality? and 3) How efficiency and effectiveness are balanced?

4.1 Experimental Setup

Dataset. All experiments are conducted on MIMIC-CXR (Johnson et al., 2019a; 2024), a large-scale chest X-ray dataset containing radiographic studies paired with free-text radiology reports. We evaluate across six clinically significant pathologies: Pneumonia, Effusion, Consolidation, Edema, Cardiomegaly, and Atelectasis. For reliable evaluation, we use a radiologist-verified evaluation subset, MIMIC-500, consisting of 482 cases with confirmed pathology labels agreed upon by CheXpert (Chambon et al., 2024) and NegBio labelers, as test queries, while the remaining as the database corpus.

Evaluation protocol. Each query consists of a patient's chest X-ray and its associated radiology report under a specific diagnostic context (e.g., "Does this patient have Pneumonia?"). The system retrieves the top-k most similar cases from the corpus. We report all metrics at $k \in \{1, 3, 5, 10\}$. Moreover, **macro-averaged** is computed across all queries regardless of finding type, reflecting overall system performance. Appendix also includes separated analysis on positive (pathology is present) and negative (normal) cases.

Table 1: Macro-averaged retrieval performance across six pathologies. **Stage 1**: standalone encoder baselines. **Stage 2**: re-ranking using the Stage 1 retriever as the candidate generator. Best per column in **bold**; second-best underlined. (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens. Complete results for each can be found in the Appendix.

	Recall@k				**Precision@k**				**MRR@k**			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Stage 1: Pretrained encoders												
BiomedCLIP (I)	.612	.786	.846	.902	.612	.583	.575	.562	.612	.673	.699	.711
BiomedCLIP (T)	.668	.829	.869	.916	.668	.617	.594	.582	.668	.721	.745	.755
MedCLIP-ResNet (T)	.619	.788	.848	.904	.619	.589	.578	.562	.619	.678	.703	.718
MedCLIP-ViT (T)	.624	.780	.836	.912	.624	.583	.573	.561	.624	.682	.708	.722
MedSigLIP (I)	.667	.826	.873	.938	.667	.624	.610	.596	.667	.720	.744	.758
MedSigLIP (T)	.666	.809	.858	.927	.666	.620	.598	.584	.666	.714	.737	.750
XRaySigLIP (I)	.669	.823	.874	.923	.669	.617	.601	.587	.669	.719	.744	.756
XRaySigLIP (T)	.633	.792	.846	.913	.633	.578	.564	.543	.633	.684	.711	.724
CheXficient (I)	.675	.827	.883	.924	.675	.620	.607	.592	.675	.742	.755	.760
CheXficient (T)	.713	.863	.900	.944	.713	.675	.656	.639	.713	.780	.789	.795
CheXficient (A)	.714	.860	.904	.938	.714	.671	.652	.630	.714	.780	.791	.795
RadIR (I)	.703	.849	.900	.945	.703	.643	.634	.615	.703	.750	.774	.785
Stage 2: Re-ranking												
RadIR→ CheXagent (T)	.702	.850	.893	.942	.702	.648	.631	.615	.702	.767	.777	.783
RadIR→ CheXagent (I)	.704	.850	.903	.943	.704	.642	.632	.616	.704	.768	.780	.785
RadIR→ CheXagent (A)	**.760**	**.893**	**.926**	**.962**	**.760**	**.708**	**.682**	**.649**	**.759**	**.819**	**.826**	**.831**

Metrics. We report three complementary label-based retrieval metrics. Recall@k (equivalently, Hit Rate@k) is a binary indicator of whether at least one relevant case appears in the top k results. Precision@k measures the fraction of top-k results that are relevant. MRR@k (Mean Reciprocal Rank) captures how quickly the first relevant result is surfaced, rewarding systems that place relevant cases at the top of the ranked list. This is vital in clinical workflows where the highest-ranked cases will be reviewed first.

Stage 1 baselines. We compare the following pretrained medical vision-language encoders as Stage 1 retrievers using cosine similarity over FAISS-indexed embeddings (Douze et al., 2024): BiomedCLIP (Zhang et al., 2024), MedCLIP (ResNet and ViT backbones) (Wang et al., 2022), MedSigLIP (Sellergren et al., 2025), XRaySigLIP (Chen et al., 2024a), RadIR (Zhang et al., 2025), and CheXficient (Wang et al., 2026). For each model, we evaluate both image and text retrieval modalities where applicable; CheXficient is additionally evaluated under a joint image+text input setting (A) as state-of-the-art in the current literature, to our best knowledge.

Stage 2 configurations. We evaluate Stage 2 re-ranking configurations using CheXagent (Chen et al., 2024b) (state-of-the-art multimodal medical LLM) token embeddings processed through the token-set to token-set similarity algorithm (T=30 iterations sufficient to converge empirically, τ=0.01 empirically for multimodal). 3.2 describes the default token extraction for the **(A)** configuration (image + text input, filtered to context text tokens). We additionally evaluate two partial configurations to isolate the contribution of each modality. In the **(T)** text-only setting, CheXagent receives only the report text and context prompt (no image); the same L_c context-question positions are extracted, yielding $T_q, T_{z_i} \in \mathbb{R}^{L_c \times d}$, but these tokens lack any visual signal. In the **(I)** image + text, unfiltered setting, CheXagent receives all three inputs jointly; however, instead of filtering to context text positions, we extract all L_{img} image patch token positions from the final decoder layer, yielding $T_q, T_{z_i} \in \mathbb{R}^{L_{\text{img}} \times d}$.

Implementation details. Stage 1 retrieval uses FAISS for approximate nearest-neighbor search over precomputed embeddings. For Stage 2, we extract token-level embeddings from CheXagent's



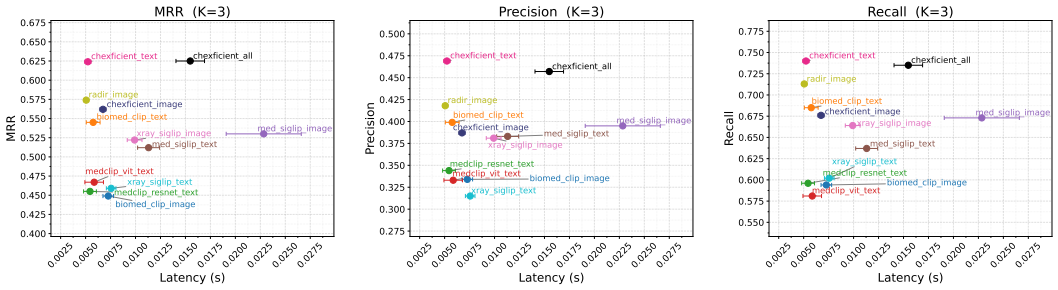
Figure 2: Top-5 candidates in each row provided by each method for one randomly sampled positive Effusion query on the top.

jointly trained vision-text encoder and compute pairwise dot-product similarity matrices, which are then processed by the token-set to token-set similarity algorithm. To run the experiments, we use two NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPUs along with a 36-core Intel(R) Xeon(R) w9-3475X. Unless otherwise noted, we use m=50 as determined by the ablation studies in Appendix C.

4.2 Performance Comparison

Table 1 presents macro-averaged recall, precision, and MRR across all Stage 1 baselines and Stage 2 configurations using RadIR as the candidate generator.

Stage 1 performance analysis. Among Stage 1 baselines, CheXficient achieves the strongest overall performance: CheXficient (A) obtains the highest Recall@1, Precision@1, and MRR@1 among Stage 1 models, while CheXficient (T) attains the highest Recall@3 and leads on Precision@3–10 and MRR@10. RadIR (I) remains competitive and is the sole Stage 1 model to lead on Recall@10 (0.945). Among the contrastive encoders, MedSigLIP and XRaySigLIP demonstrate competitive



(a) MRR @ K=3 (b) Precision @ K=3 (c) Recall @ K=3

Figure 3: Stage 1 average retrieval performance vs. inference latency at $k = 3$ on local hardware. Horizontal bars denote latency std (cutoff-independent).

Table 2: Stage 2 average re-ranking latency with different candidate pool size m.

m	10	25	50	100	200
time±std (s)	2.284±.144	5.336±.231	9.419±.251	19.408±.386	39.830±4.831

performance, particularly in image-modality retrieval. Text-modality retrieval generally matches or exceeds image-modality retrieval for BiomedCLIP and MedCLIP, but the pattern reverses for MedSigLIP and XRaySigLIP, indicating that modality preference is model-dependent.

To determine whether a stronger Stage 1 backbone also yields better Stage 2 candidate pools, we compare RadIR (I) against CheXficient (T) and (A) as alternative candidate generators for re-ranking in Appendix Section B, where RadIR shows the advantage in most cases.

Stage 2 performance analysis. Within the token-modality comparison (Table 1), CheXagent (A) achieves the best performance across all three input settings, with particularly pronounced gains in ranking quality: MRR@1 reaches 0.759 (+5.6 pp over RadIR (I)), MRR@3 reaches 0.819 (+6.9 pp), and MRR@10 reaches 0.831 (+4.6 pp), indicating that the re-ranker not only retrieves more relevant cases but places them at higher positions in the ranked list. Recall@1 improves to 0.760 and Precision@3 to 0.708 (+6.5 pp over RadIR (I)). More importantly, stage 2 consistently outperforms all stage 1 baselines significantly, which can also be observed in the example illustration in Fig. 2

Efficiency analysis. Fig. 3 shows the performance vs. efficiency analysis for Stage 1 using $k = 3$ as an example, while Table 2 shows the latency of Stage 2 with different candidate pool size m. We can observe that vector database based similarity search is much faster than the token-set to token-set re-rank as expected. Therefore, we set $m = 50$ and this sacrifice of efficiency is desirable considering the above demonstrated performance boost. More ablation studies or details can be found in the Appendix.

5 Conclusion

To perform a context-aware retrieval of similar past patients using both image and text data without any additional training, we propose a two-stage approach. To enhance the efficiency of retrieval from a large corpus, we take advantage of existing vector database based similarity search to generate a smaller pool of candidates in the first stage. Thereafter, to improve retrieval performance considering both image and text data, we propose to take advantage of pre-trained multimodal medical LLMs by capturing context and fine-grained semantics through token-set to token-set similarity without further training, while existing methods often require additional training to address this need. Our comprehensive experiments demonstrate both the efficiency and efficacy of our proposal. More quantitative and qualitative evaluations on more diverse medical data with clinical validations from clinicians will be an interesting future work.

AI USE STATEMENT

In this work, we have used generative AI tools to provide feedback on research methodology or experiments. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design research methodology or experiments, or support qualitative and thematic data analysis. Generating synthetic datasets, cleaning and reformatting datasets, and assisting with translation are not applicable to this work.

Additionally, we used generative AI tools to support code debugging, summarize or analyze existing literature, sourcing/searching for information, improve paper readability, and format references. We have reviewed all AI-assisted work by reading and filtering related research, performing code reviews via GitHub issues and pull requests (both manual and AI-assisted), and proofreading the research paper. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Naren Akash R. J., Arihanth Tadanki, and Jayanthi Sivaswamy. Chextriev: Anatomy-centered representation for case-based retrieval of chest radiographs. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part I*, pp. 461–471, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72377-3. doi: 10.1007/978-3-031-72378-0_43. URL https://doi.org/10.1007/978-3-031-72378-0_43.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014. ISBN 978-0-521-83378-3. doi: 10.1017/CBO9780511804441. URL https://web.stanford.edu/%7Eboyd/cvxbook/.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats. Stanford AIMI Shared Dataset, 2024. URL https://stanfordaimi.azurewebsites.net/datasets/5158c524-d3ab-4e02-96e9-6ee9efc110a1.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net, 2023. URL https://openreview.net/forum?id=zqwryBoXYnh.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment, 2020. URL https://arxiv.org/abs/2006.14744.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Andrew Johnston, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024a.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b. URL https://arxiv.org/abs/2401.12208.

Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2502.15040.

Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. Infoot: Information maximizing optimal transport, 2023. URL https://arxiv.org/abs/2210.03164.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. URL https://arxiv.org/abs/1306.0895.

Le Dong, Qixuan Cao, Chunlei Li, Jingliang Hu, Yilei Shi, Weisheng Dong, Xiao Xiang Zhu, and Lichao Mou. Rethinking contrastive language-image pre-training for medical cross-modal retrieval: Beyond one-to-one correspondence, 2025. URL https://openreview.net/forum?id=HPAzsibJFM.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Jiansheng Fang, Huazhu Fu, and Jiang Liu. Deep triplet hashing network for case-based medical image retrieval. *Medical Image Analysis*, 69:101981, 2021. doi: 10.1016/j.media.2020.101981.

Shichao Fang, Shenda Hong, Qing Li, Pengfei Li, Tim Coats, Beiji Zou, and Guilan Kong. Cross-modal similar clinical case retrieval using a modular model based on contrastive learning and k-nearest neighbor search. *International Journal of Medical Informatics*, 193:105680, 2025. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2024.105680. URL https://www.sciencedirect.com/science/article/pii/S1386505624003435.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Zhiqi Huang, Puxuan Yu, and James Allan. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pp. 1048–1056. ACM, February 2023. doi: 10.1145/3539597.3570468. URL http://dx.doi.org/10.1145/3539597.3570468.

Intelerad. PACS in radiology, January 2022. URL https://www.intelerad.com/en/2022/01/24/radiology-pacs-system/. Accessed: 2026-01-28.

Alistair Johnson et al. MIMIC-CXR Database, 2024. URL https://doi.org/10.13026/4jqj-jw95. RRID:SCR_007345.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019a. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

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

Farnaz Khun Jush, Steffen Vogler, and Matthias Lenga. Content-based 3D image retrieval and a ColBERT-inspired re-ranking for tumor flagging and staging. *Journal of Imaging Informatics in Medicine*, 2025. doi: 10.1007/s10278-025-01598-0. URL https://doi.org/10.1007/s10278-025-01598-0.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Ashnil Kumar, Jinman Kim, Weidong Cai, Michael Fulham, and Dagan Feng. Content-based medical image retrieval: a survey of applications to multidimensional and multimodality data. *Journal of Digital Imaging*, 26(6):1025–1039, December 2013. doi: 10.1007/s10278-013-9619-2.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Yunhao Liu, Suyang Xi, Shiqi Liu, Hong Ding, Chicheng Jin, Chong Zhong, Junjun He, Catherine C. Liu, and Yiqing Shen. Multimodal medical image binding via shared text embeddings, 2025. URL https://arxiv.org/abs/2506.18072.

Patrick Matulich. RadLIT: Late interaction retrieval for radiology education, 2026. URL https://huggingface.co/datasets/matulichpt/radlit9-benchmark. Models and benchmark available at https://huggingface.co/matulichpt.

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. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Yossi Rubner, Carlo Tomasi, and Leonidas Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 11 2000. doi: 10.1023/A:1026543900054.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Mai A. Shaaban, Tausifa Jan Saleem, Vijay Ram Kumar Papineni, and Mohammad Yaqub. MOTOR: Multimodal optimal transport via grounded retrieval in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer, 2025. URL https://github.com/BioMedIA-MBZUAI/MOTOR.

Wilson Silva, Tiago Gonçalves, Kirsi Härmä, Erich Schröder, Verena Carola Obmann, María Cecilia Barroso, Alexander Poellinger, Mauricio Reyes, and Jaime S. Cardoso. Computer-aided diagnosis through medical image retrieval in radiology. *Scientific Reports*, 12(1):20732, December 2022. doi: 10.1038/s41598-022-25027-2. URL https://doi.org/10.1038/s41598-022-25027-2.

Steven Song, Anirudh Subramanyam, Irene Madejski, and Robert L. Grossman. Lab-rag: Label boosted retrieval augmented generation for radiology report generation, 2025. URL https://arxiv.org/abs/2411.16523.

Chong Wang, Yabin Zhang, Yunhe Gao, Maya Varma, Clemence Mottez, Faidra Patsatzi, Jiaming Liu, Jin Long, Jean-Benoit Delbrouck, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. A data- and compute-efficient chest X-ray foundation model beyond aggressive scaling. https://arxiv.org/abs/2602.22843v1, February 2026.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022. URL https://arxiv.org/abs/2210.10163.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest ImaGenome Dataset. *PhysioNet*, July 2021. doi: 10.13026/wv01-y230. URL https://doi.org/10.13026/wv01-y230. Version 1.0.0.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models, 2025. URL https://arxiv.org/abs/2410.13085.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders, 2023. URL https://arxiv.org/abs/2308.01317.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

Tengfei Zhang, Ziheng Zhao, Chaoyi Wu, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radir: A scalable framework for multi-grained medical image retrieval via radiology report mining, 2025. URL https://arxiv.org/abs/2503.04653.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation, 2024. URL https://arxiv.org/abs/2406.16845.

A Component Ablation Study

Table 3: Macro-averaged retrieval performance across six pathologies. Best per column in **bold**; second-best underlined. (I)=image+text input, unfiltered; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Precision@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
RadIR (I)→ CheXagent (A) (L0)	.596	.794	.854	.941	.596	.598	.602	.614	.596	.684	.698	.710
RadIR (I)→ CheXagent (A) (Ln)	.717	.870	.914	**.963**	.717	.662	.631	.610	.717	.786	.796	.803
RadIR (I)→ CheXagent (A) (LnSec)	.719	.865	.913	.954	.719	.663	.633	.611	.719	.785	.796	.801
RadIR (I)→ CheXagent (A) (LnFilt)	.747	**.893**	**.929**	.952	.747	.706	.676	.647	.747	.813	.821	.824
RadIR (I)→ CheXagent (A) (LnSecFilt)	**.760**	**.893**	.926	.962	**.760**	**.708**	**.682**	**.649**	**.759**	**.819**	**.826**	**.831**

Component ablation analysis. Our ablation study in Table 3 isolates each component's contribution to the full LnSecFilt configuration. Removing the label context entirely (L0) reduces Recall@1 and MRR@1 to 0.596, placing this configuration below every Stage 1 baseline in Table 1. Adding the label context alone (Ln) produces the largest single recovery: Recall@1 rises to 0.717 and MRR@10 to 0.803, surpassing both RadIR (I) (0.703) and CheXficient (A) (0.714) on Recall@1. Appending the visual token filter (LnFilt) yields the next-largest gain in Recall@1 of 0.747 (+3.0 pp over Ln), Recall@3 of 0.893, and Recall@5 of 0.929. By contrast, adding section context to the label alone (LnSec) contributes only +0.2 pp on Recall@1 and reduces Recall@10 from 0.963 to 0.954 relative to Ln.

The full LnSecFilt configuration achieves the best Recall@1 (0.760), all Precision metrics, and all MRR values in Table 3. LnSecFilt and LnFilt are the two strongest configurations overall, with the section context prompt being the primary differentiating factor between them: LnFilt outperforms LnSecFilt on Recall@5 (0.929 vs. 0.926) and ties on Recall@3 (0.893), while LnSecFilt leads on Recall@1, all Precision, and all MRR values. Ln leads uniquely on Recall@10 at 0.963, only slightly above LnSecFilt (0.962). The most critical comparison in this ablation is between configurations that include the context-question filter (LnFilt and LnSecFilt) and those that do not (Ln and LnSec): removing Filt consistently reduces Recall@1 by 3.0–4.1 pp, with the full LnSecFilt configuration outperforming LnSec by 4.1 pp on Recall@1 (0.760 vs. 0.719).

B Ablation Study: CheXficient vs. RadIR

Table 4: Macro-averaged retrieval performance across six pathologies. Best per column in **bold**; second-best underlined. (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Precision@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
CheXficient (T)→ CheXagent (A)	.748	.891	**.931**	.950	.748	.700	.681	**.656**	.748	.813	.822	.824
CheXficient (A)→ CheXagent (A)	.755	.889	.928	.949	.755	.701	.675	.650	.755	.815	.824	.827
RadIR (I)→ CheXagent (A)	**.760**	**.893**	.926	**.962**	**.760**	**.708**	**.682**	.649	**.759**	**.819**	**.826**	**.831**

As seen in Table 1, CheXficient performs better than RadIR during Stage 1; however, upon running our ablation study, we see that RadIR performs better, and acknowledge these results seem counterintuitive. The superior Stage 2 performance of RadIR (I) over CheXficient (T) and CheXficient (A) as a candidate generator can be attributed primarily to differences in training objectives. CheXficient is optimized with an InfoNCE contrastive loss that frames retrieval as a binary pairing task. This produces embeddings suited to Stage 1, but does not explicitly shape the embedding space to reflect graded clinical similarity among candidates. In contrast, RadIR jointly optimizes a masked InfoNCE loss and a triplet loss against a continuous image-image similarity matrix derived from

RaTEScore (Zhang et al., 2025). This triplet objective directly encodes graded relevance ordering into the embedding space, meaning the m=50 candidate pool it forwards to Stage 2 contains more meaningful ordinal structure for the re-ranker to exploit.

A secondary factor is the modality basis of pool selection. CheXficient (T) selects candidates via text embedding similarity, which may surface cases whose reports are semantically close but whose visual content diverges from the query image. Since Stage 2 re-ranking operates over CheXagent token representations extracted from joint image–text encoding, a pool composed of visually coherent candidates provides a better area for token-level discrimination. Together, these differences suggest that retrieval-specialized training objectives, rather than raw Stage 1 embedding quality, are the more important factor in determining Stage 2 effectiveness.

C Ablation Study: Candidate Pool Size m

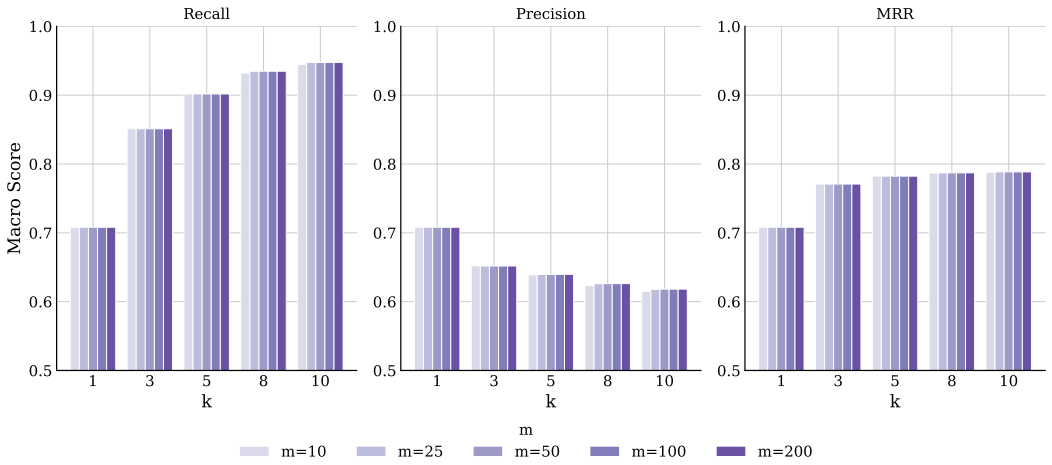


Figure 4: Macro-averaged retrieval performance across pathologies as a function of retrieval depth (k) for varying m configurations. Each subplot reports a different ranking metric — Recall, Precision, and Mean Reciprocal Rank (MRR) — averaged over six pathology classes (Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Pneumonia). Bars are grouped by k value and colored by m.

To determine the optimal number of Stage 1 candidates forwarded to Stage 2, we evaluate pool sizes $m \in \{10, 25, 50, 100, 200\}$ on the MIMIC-500. Results show that performance is fairly robust across pool sizes. When $m \in \{10, 25, 200\}$, we observe a small drop in performance, while $m \in \{50, 100\}$ achieve equivalent results. We select m=50 as it provides the same retrieval quality as m=100 while reducing the computational cost of Stage 2.

D Ablation Study: Interpolation Parameter λ

To measure the influence of the token embeddings for re-ranking, we perform an ablation study using an inverse-weighting parameter λ. To do this, we combine the scores in Stage 1 and Stage 2 via linear interpolation: $s_i^{(2)} = (1-\lambda) \cdot s_i^{(1)} + \lambda \cdot w_i$. We investigate the effect of λ on retrieval quality by testing $\lambda \in \{0.0, 0.1, 0.3, 0.5, 0.8, 1.0\}$ with m=50 on the MIMIC-500 validation set. Note that λ=0.0 corresponds to Stage 1 retrieval only (no re-ranking), while λ=1.0 relies entirely on the token-set to token-set similarity. We find that λ=1.0 achieves the best performance overall, followed by λ=0.8. Performance improves monotonically as more weight is placed on our proposal, indicating that the token-set to token-set similarity captures retrieval-relevant information that is not fully represented in the Stage 1 cosine similarities.

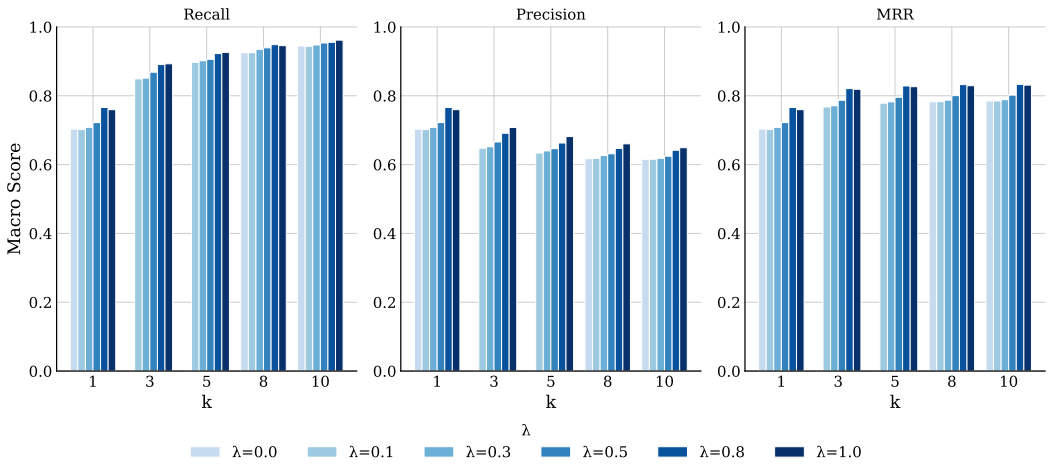


Figure 5: Macro-averaged retrieval performance across pathologies as a function of retrieval depth (k) for varying interpolation weights (λ).

E Stage 1 Performance vs Time

Fig. 6 plots retrieval performance (recall, precision, and MRR at $k \in \{1, 3, 5\}$) against per-query latency for all Stage 1 encoders, where latency is the mean $\pm$ standard deviation of per-query search time over the 482 test queries, excluding offline index construction. All Stage 1 encoders run in under 25 ms per query, and latency is largely uncorrelated with retrieval quality: RadIR (I) and CheXficient (T) are simultaneously the fastest and among the most accurate encoders, whereas the slowest encoder, MedSigLIP (I) provides no corresponding accuracy gain. Fusing CheXficient's image and text embeddings roughly triples latency without improving over its text-only variant.

F Per-Pathology Results Discussion

This section accompanies the seven per-pathology results tables (Tables 5 to 11), each of which reports Recall@k, Precision@k, and MRR@k at $k \in \{1, 3, 5, 10\}$ across macro-averaged, positive-class, and negative-class subsets.

Macro-averaged (Table 5). Among Stage 1 retrievers, CheXficient achieves the strongest overall performance: CheXficient (A) leads on Recall@1 (.714), Precision@1 (.714), and MRR@1 (.714), while CheXficient (T) attains the highest Recall@3 (.863), Precision@3–10, and MRR@10 (.795). RadIR (I) remains competitive and leads on Recall@10 (.945). Our best Stage 2 variant, RadIR (I)$\rightarrow$ CheXagent (A), lifts macro Recall@1 from .714 (CheXficient (A)) to .760 (+4.6 pp) and positive Recall@1 from .533 (CheXficient (A)) to .612 (+7.9 pp), with the largest absolute gains concentrated on the positive class. This shows that token-level reranking recovers fine-grained pathology-relevant features that global embeddings are unable to capture. Negative-class retrieval is already near saturation in Stage 1 (CheXficient (T) .896 Recall@1), so Stage 2 gains there are small but uniformly positive.

Pneumonia (Table 6). Stage 1 is led by CheXficient (A) (.705 Recall@1) rather than RadIR (I) (.642), with CheXficient (T) close behind at .701. Both CheXficient variants outperform MedSigLIP (T) (.684). CheXficient (A) also leads on positive Recall@1 (.433) and achieves the highest MRR@10 (.778) among Stage 1 models. Our CheXagent (A) Stage 2 variant extends this to .719 Recall@1 (+1.4 pp over CheXficient (A)) and raises positive Recall@1 to .467 (+3.4 pp), but the single-modality variants do not consistently beat the CheXficient baselines.

Effusion (Table 7). Effusion is the easiest of the six pathologies, with RadIR (I) leading Stage 1 on macro Recall@1 (.830) and Recall@10 (.994). CheXficient (T) is competitive on positive Recall@1 (.819, the highest among Stage 1 models). Stage 2 contributes ranking quality more than raw retrieval

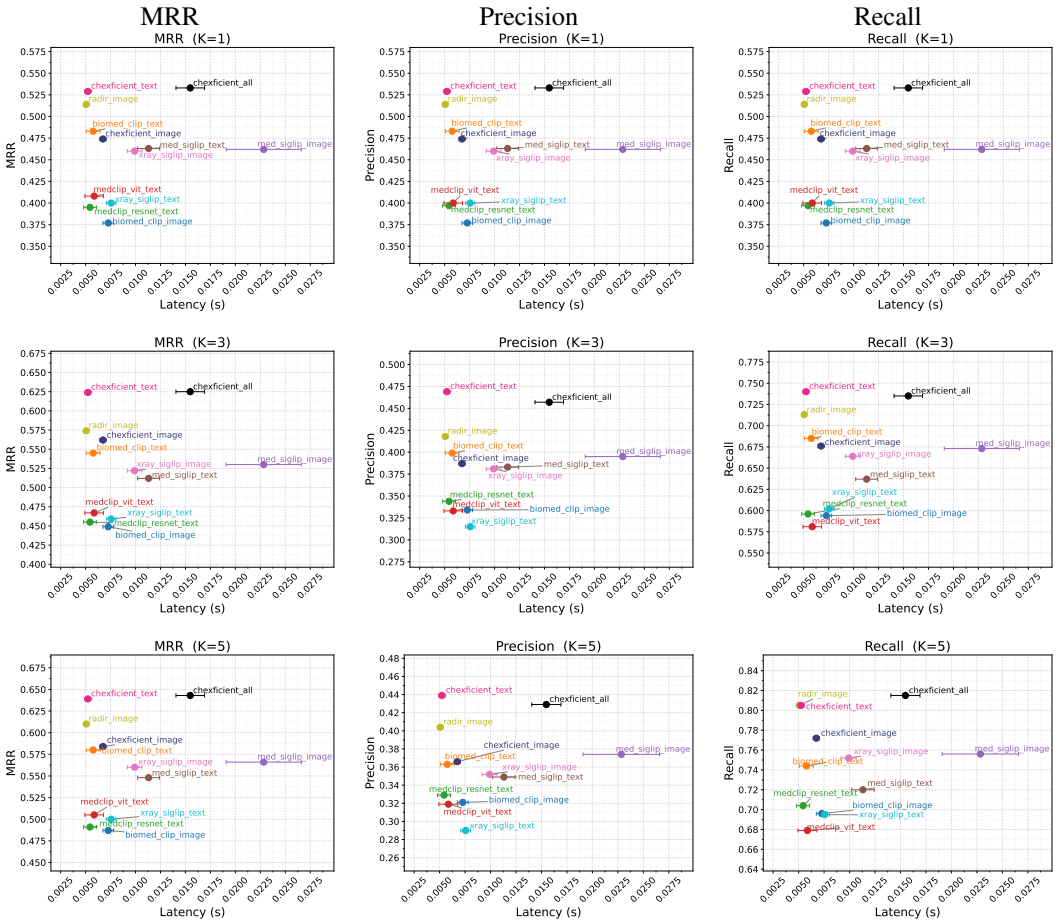


Figure 6: Retrieval performance vs. inference latency across metrics MRR, Precision, Recall and $k = 1, 3, 5$). Horizontal bars denote latency std.

breadth. CheXagent (A) achieves perfect Recall (1.000) at $k = 10$ across all three subsets and a $+1.0$-point gain in positive Precision@1. The narrow Stage 2 margin here is structural rather than a failure mode: there is little headroom left to recover above a ceiling-bound Stage 1.

Consolidation (Table 8). Consolidation is the hardest pathology: most Stage 1 models do not perform well on the positive class, with BiomedCLIP (I) being the worst. CheXficient (A) leads Stage 1 on macro Recall@1 (.600) and MRR@1 (.600), while RadIR (I) leads on Recall@10 (.860). Our CheXagent (A) Stage 2 variant produces the largest absolute improvements observed in any table. Positive Recall@3 jumps from .240 (RadIR (I)) to .680 (+44 pp), positive Recall@1 from .240 (CheXficient (A)) to .280 (+4 pp), and macro Recall@3 from .700 (BiomedCLIP (T), best Stage 1 at k=3) to .839 (+13.9 pp).

Edema (Table 9). CheXficient (A) leads Stage 1 on macro Recall@1 (.800), positive Recall@1 (.697), and MRR@10 (.871), surpassing both MedSigLIP (I) and XRaySigLIP (I) (.771 Recall@1) that previously led without CheXficient. CheXagent (A) pushes macro Recall@1 to .862 (+6.2 pp over CheXficient (A)) and positive Recall@1 to .818 (+12.1 pp), and is the best variant in every column of every block.

Cardiomegaly (Table 10). Stage 1 is led by RadIR (I) (.776 Recall@1, .673 positive Recall@1), with CheXficient (T) close behind (.754 Recall@1, .626 positive Recall@1), along with BiomedCLIP (T) (.719 Recall@1). CheXagent (A) lifts macro Recall@1 to .801 (+2.5 pp over RadIR (I)) and positive Recall@1 to .701 (+2.8 pp). Notably, all three Stage 2 variants beat the strongest Stage 1

baseline on positive-class metrics, suggesting that for pathologies with clear, consistently-described anatomical correlates, the benefit of token-level reranking is robust to the choice of reranker modality.

Atelectasis (Table 11). CheXficient (T) leads Stage 1 on macro Recall@1 (.671) and MRR@10 (.789), with RadIR (I) close behind (.661 Recall@1). The heterogeneous visual presentation of atelectasis is reflected in tight clustering across Stage 1 models (.62–.67 Recall@1) and in XRaySigLIP (T)'s strong negative-class performance (.825 Recall@1). CheXagent (A) raises macro Recall@1 to .737 (+6.6 pp over CheXficient (T)) and positive Recall@3 from .676 (RadIR (I)) to .827 (+15.1 pp).

Three patterns recur across all seven tables and are worth surfacing in the main paper. First, RadIR (I)→CheXagent (A) wins or ties on macro Recall@1 in every table, surpassing both RadIR (I) and the CheXficient variants despite the latter achieving stronger Stage 1 performance on most metrics. Second, the gap between CheXagent (A) and the best Stage 1 baseline scales with pathology difficulty: it is smallest on Effusion and Cardiomegaly (where Stage 1 performance is already strong) and largest on Consolidation and Edema (where Stage 1 struggles most on the positive class). Third, negative-class retrieval is at or near ceiling across all pathologies, which we acknowledge is partly an artifact of the binary same-pathology-label relevance criterion: under this criterion, "any non-pathology case" is a very loose match, and the resulting saturation limits the discriminative power of negative-class metrics.

Table 5: Retrieval performance on **Macro (A)veraged (All Six Pathologies)** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.612	.786	.846	.902	.612	.583	.575	.562	.612	.673	.699	.711
BiomedCLIP (T)	.668	.829	.869	.916	.668	.617	.594	.582	.668	.721	.745	.755
MedCLIP-ResNet (T)	.619	.788	.848	.904	.625	.589	.578	.562	.625	.678	.703	.718
MedCLIP-ViT (T)	.624	.780	.836	.912	.624	.583	.573	.561	.628	.682	.708	.722
MedSigLIP (I)	.667	.826	.873	.938	.667	.624	.610	.596	.667	.720	.744	.758
MedSigLIP (T)	.666	.809	.858	.927	.666	.620	.598	.584	.666	.714	.737	.750
XRaySigLIP (I)	.669	.823	.874	.923	.669	.617	.601	.587	.669	.719	.744	.756
XRaySigLIP (T)	.633	.792	.846	.913	.633	.578	.564	.543	.633	.684	.711	.724
CheXficient (I)	.675	.827	.883	.924	.675	.620	.607	.592	.675	.742	.755	.760
CheXficient (T)	.713	.863	.900	.944	.713	.675	.656	.639	.713	.780	.789	.795
CheXficient (A)	.714	.860	.904	.938	.714	.671	.652	.630	.714	.780	.791	.795
RadIR (I)	.703	.849	.900	.945	.703	.643	.634	.615	.703	.750	.774	.785
Ours: RadIR (I)→CXA (T)	.702	.850	.893	.942	.702	.648	.631	.615	.702	.767	.777	.783
Ours: RadIR (I)→CXA (I)	.704	.850	.903	.943	.704	.642	.632	.616	.704	.768	.780	.785
Ours: RadIR (I)→CXA (A)	**.760**	**.893**	**.926**	**.962**	**.760**	**.708**	**.682**	**.649**	**.759**	**.819**	**.826**	**.831**
Positive Findings												
BiomedCLIP (I)	.377	.594	.696	.804	.377	.334	.321	.304	.377	.449	.487	.511
BiomedCLIP (T)	.483	.685	.744	.861	.483	.399	.363	.347	.483	.545	.580	.597
MedCLIP-ResNet (T)	.397	.596	.704	.809	.397	.344	.329	.302	.395	.455	.491	.520
MedCLIP-ViT (T)	.400	.581	.679	.825	.400	.333	.319	.301	.408	.467	.505	.531
MedSigLIP (I)	.462	.673	.756	.876	.462	.395	.374	.357	.462	.530	.566	.591
MedSigLIP (T)	.463	.637	.720	.853	.463	.383	.349	.328	.463	.512	.548	.574
XRaySigLIP (I)	.460	.664	.752	.847	.460	.381	.352	.336	.460	.522	.560	.582
XRaySigLIP (T)	.400	.602	.695	.826	.400	.315	.290	.262	.400	.459	.500	.525
CheXficient (I)	.474	.676	.772	.848	.474	.387	.366	.340	.474	.562	.584	.594
CheXficient (T)	.529	.740	.805	.889	.529	.469	.439	.413	.529	.624	.639	.651
CheXficient (A)	.533	.735	.815	.879	.533	.457	.429	.393	.533	.625	.643	.652
RadIR (I)	.514	.713	.805	.891	.514	.418	.404	.375	.514	.574	.610	.632
Ours: RadIR (I)→CXA (T)	.504	.713	.791	.885	.504	.422	.397	.372	.504	.593	.611	.623
Ours: RadIR (I)→CXA (I)	.517	.715	.810	.889	.517	.417	.403	.375	.517	.602	.623	.634
Ours: RadIR (I)→CXA (A)	**.612**	**.803**	**.858**	**.923**	**.612**	**.532**	**.486**	**.427**	**.612**	**.696**	**.708**	**.716**
Negative Findings												
BiomedCLIP (I)	.848	.977	**.997**	.999	.848	.833	.829	.821	.848	.897	.911	.912
BiomedCLIP (T)	.853	.973	.993	.999	.853	.836	.826	.816	.853	.897	.911	.913
MedCLIP-ResNet (T)	.848	.981	.993	**1.000**	.848	.834	.827	.822	.856	.901	.914	.916
MedCLIP-ViT (T)	.848	.980	.993	.999	.848	.833	.827	.821	.849	.897	.911	.913
MedSigLIP (I)	.872	.979	.991	.999	.872	.853	.846	.836	.872	.911	.922	.924
MedSigLIP (T)	.869	.981	.996	**1.000**	.869	.856	.848	.841	.870	.916	.925	.926
XRaySigLIP (I)	.878	.982	.996	.999	.878	.853	.849	.839	.878	.917	.928	.929
XRaySigLIP (T)	.866	.981	**.997**	**1.000**	.866	.841	.837	.823	.866	.910	.921	.923
CheXficient (I)	.874	.979	.993	**1.000**	.874	.852	.848	.844	.874	.922	.925	.926
CheXficient (T)	.896	.985	.996	.998	.896	.881	.873	.865	.896	.937	.939	.940
CheXficient (A)	.894	.985	.993	.998	.894	**.885**	.876	.868	.894	.936	.938	.939
RadIR (I)	.892	.985	.996	.998	.892	.868	.864	.856	.892	.927	.937	.938
Ours: RadIR (I)→CXA (T)	.901	**.987**	.995	.999	.901	.874	.866	.860	.901	.940	.942	.943
Ours: RadIR (I)→CXA (I)	.891	.985	.996	.998	.891	.867	.861	.857	.891	.934	.937	.937
Ours: RadIR (I)→CXA (A)	**.907**	.983	.994	**1.000**	**.907**	.883	**.877**	**.871**	**.907**	**.942**	**.945**	**.946**

Table 6: Retrieval performance on **Pneumonia** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.611	.683	.750	.833	.611	.547	.553	.535	.611	.629	.645	.665
BiomedCLIP (T)	.613	.683	.700	.750	.613	.539	.516	.509	.614	.639	.649	.656
MedCLIP-ResNet (T)	.598	.667	.717	.783	.615	.533	.531	.524	.598	.623	.635	.659
MedCLIP-ViT (T)	.631	.700	.717	.817	.631	.554	.531	.521	.618	.641	.647	.670
MedSigLIP (I)	.644	.733	.800	.900	.644	.598	.578	.552	.645	.671	.687	.712
MedSigLIP (T)	.684	.766	.783	.833	.684	.574	.538	.528	.684	.715	.726	.732
XRaySigLIP (I)	.638	.733	.783	.883	.638	.575	.551	.534	.638	.669	.684	.704
XRaySigLIP (T)	.650	.733	.750	.850	.650	.556	.529	.519	.650	.657	.680	.696
CheXficient (I)	.668	.732	.833	.867	.668	.569	.565	.541	.668	.700	.721	.725
CheXficient (T)	.701	.800	.833	.900	.701	.611	.591	.560	.701	.745	.751	.760
CheXficient (A)	.705	**.817**	**.867**	**.933**	.705	.613	.588	.553	.705	**.758**	**.769**	**.778**
RadIR (I)	.642	.800	.833	.850	.642	.581	.561	.527	.642	.704	.720	.725
Ours: RadIR (I)→CXA (T)	.610	.783	.817	.850	.610	.575	.559	.527	.610	.688	.697	.701
Ours: RadIR (I)→CXA (I)	.659	.800	.833	.867	.659	.578	.561	.528	.659	.718	.726	.729
Ours: RadIR (I)→CXA (A)	**.719**	.767	.817	.883	**.719**	**.622**	**.592**	**.562**	**.719**	.740	.750	.759
Positive Findings												
BiomedCLIP (I)	.267	.367	.500	.667	.267	.144	.153	.120	.267	.283	.314	.353
BiomedCLIP (T)	.267	.367	.400	.667	.267	.122	.080	.070	.267	.300	.319	.332
MedCLIP-ResNet (T)	.267	.333	.433	.567	.267	.111	.113	.103	.233	.267	.289	.338
MedCLIP-ViT (T)	.300	.400	.433	.633	.300	.156	.113	.097	.267	.300	.311	.357
MedSigLIP (I)	.333	.467	.600	.800	.333	.244	.213	.167	.333	.367	.397	.447
MedSigLIP (T)	.400	.533	.567	.667	.400	.200	.140	.107	.400	.450	.469	.481
XRaySigLIP (I)	.300	.467	.567	.767	.300	.189	.140	.110	.300	.350	.381	.421
XRaySigLIP (T)	.333	.467	.500	.700	.333	.156	.100	.087	.333	.333	.378	.409
CheXficient (I)	.367	.467	.667	.733	.367	.189	.180	.137	.367	.417	.458	.467
CheXficient (T)	.433	.600	.667	.800	.433	.267	**.227**	**.170**	.433	.506	.519	.537
CheXficient (A)	.433	**.633**	**.733**	**.867**	.433	.267	.220	.157	.433	**.528**	**.551**	**.568**
RadIR (I)	.333	.600	.667	.700	.333	.211	.180	.110	.333	.433	.464	.475
Ours: RadIR (I)→CXA (T)	.267	.567	.633	.700	.267	.200	.173	.110	.267	.400	.417	.426
Ours: RadIR (I)→CXA (I)	.367	.600	.667	.733	.367	.211	.180	.113	.367	.461	.476	.484
Ours: RadIR (I)→CXA (A)	**.467**	.533	.633	.767	**.467**	**.289**	**.227**	.163	**.467**	.494	.516	.533
Negative Findings												
BiomedCLIP (I)	.956	**1.000**	**1.000**	**1.000**	.956	.949	.952	.951	.956	.976	.977	.977
BiomedCLIP (T)	.960	**1.000**	**1.000**	**1.000**	.960	.957	.951	.948	.960	.978	.979	.979
MedCLIP-ResNet (T)	.962	**1.000**	**1.000**	**1.000**	.962	.956	.948	.945	.962	.979	.981	.981
MedCLIP-ViT (T)	.962	**1.000**	**1.000**	**1.000**	.962	.952	.949	.945	.969	.981	.983	.983
MedSigLIP (I)	.956	**1.000**	**1.000**	**1.000**	.956	.951	.943	.936	.956	.975	.977	.977
MedSigLIP (T)	.969	.998	**1.000**	**1.000**	.969	.947	.936	.949	.969	.979	.983	.983
XRaySigLIP (I)	**.976**	**1.000**	**1.000**	**1.000**	**.976**	**.962**	**.962**	.958	**.976**	**.988**	**.988**	**.988**
XRaySigLIP (T)	.967	**1.000**	**1.000**	**1.000**	.967	.957	.958	.952	.967	.981	.983	.983
CheXficient (I)	.969	.998	**1.000**	**1.000**	.969	.948	.950	.945	.969	.983	.983	.983
CheXficient (T)	.969	**1.000**	**1.000**	**1.000**	.969	.955	.955	.950	.969	.984	.984	.984
CheXficient (A)	**.976**	**1.000**	**1.000**	**1.000**	**.976**	.960	.956	.950	**.976**	**.988**	**.988**	**.988**
RadIR (I)	.951	**1.000**	**1.000**	**1.000**	.951	.950	.943	.944	.951	.975	.975	.975
Ours: RadIR (I)→CXA (T)	.954	**1.000**	**1.000**	**1.000**	.954	.951	.945	.945	.954	.976	.976	.976
Ours: RadIR (I)→CXA (I)	.951	**1.000**	**1.000**	**1.000**	.951	.945	.942	.943	.951	.975	.975	.975
Ours: RadIR (I)→CXA (A)	.971	**1.000**	**1.000**	**1.000**	.971	.956	.958	**.960**	.971	.985	.985	.985

Table 7: Retrieval performance on **Effusion** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.664	.908	.968	.988	.664	.649	.641	.620	.664	.751	.783	.791
BiomedCLIP (T)	.731	.930	.965	.996	.731	.670	.647	.639	.731	.801	.825	.833
MedCLIP-ResNet (T)	.668	.914	.953	.992	.688	.662	.637	.614	.688	.765	.794	.802
MedCLIP-ViT (T)	.693	.904	.952	.993	.693	.652	.632	.606	.694	.771	.799	.808
MedSigLIP (I)	.752	.928	.955	.994	.752	.734	.717	.685	.752	.806	.834	.840
MedSigLIP (T)	.730	.907	.972	.997	.730	.683	.669	.657	.730	.798	.820	.830
XRaySigLIP (I)	.756	.914	.958	.991	.756	.714	.705	.692	.756	.810	.833	.842
XRaySigLIP (T)	.675	.914	.959	.995	.675	.625	.611	.587	.675	.750	.785	.794
CheXficient (I)	.727	.914	.960	.995	.727	.699	.690	.683	.727	.809	.820	.824
CheXficient (T)	.826	.943	.973	.987	.826	**.808**	**.785**	**.771**	.826	.880	.887	.889
CheXficient (A)	.819	.942	.974	.990	.819	.801	.782	.757	.819	.875	.882	.884
RadIR (I)	.830	.953	.977	.994	.830	.773	.760	.730	.830	.874	.889	.893
Ours: RadIR (I)→CXA (T)	**.836**	**.953**	.975	.996	**.836**	.775	.761	.730	**.836**	**.888**	.893	**.896**
Ours: RadIR (I)→CXA (I)	**.836**	.949	.975	.994	**.836**	.771	.759	.736	**.836**	**.888**	**.894**	**.896**
Ours: RadIR (I)→CXA (A)	.809	.951	**.985**	**1.000**	.809	.791	.774	.751	.809	.875	.883	.885
Positive Findings												
BiomedCLIP (I)	.611	.886	.953	.979	.611	.594	.594	.564	.611	.707	.744	.754
BiomedCLIP (T)	.746	.922	.964	.995	.746	.627	.603	.602	.746	.806	.830	.839
MedCLIP-ResNet (T)	.627	.876	.938	.985	.627	.625	.599	.578	.642	.728	.760	.770
MedCLIP-ViT (T)	.694	.871	.927	.990	.694	.615	.596	.565	.679	.749	.781	.791
MedSigLIP (I)	.746	.912	.948	.995	.746	.715	.695	.660	.746	.788	.822	.830
MedSigLIP (T)	.731	.865	.953	.995	.731	.649	.620	.606	.731	.775	.799	.816
XRaySigLIP (I)	.710	.865	.933	.985	.710	.663	.648	.645	.710	.759	.787	.801
XRaySigLIP (T)	.575	.881	.927	.990	.575	.549	.535	.505	.575	.676	.717	.729
CheXficient (I)	.689	.876	.948	.990	.689	.651	.637	.638	.689	.770	.787	.792
CheXficient (T)	**.819**	.917	.964	.985	**.819**	**.782**	**.756**	**.745**	**.819**	.863	.873	.876
CheXficient (A)	.808	.922	.969	.990	.808	.772	.750	.721	.808	.857	.867	.870
RadIR (I)	.798	**.943**	.964	.995	.798	.736	.719	.677	.798	.842	.864	.870
Ours: RadIR (I)→CXA (T)	.803	.938	.964	.995	.803	.732	.716	.668	.803	.861	.867	.871
Ours: RadIR (I)→CXA (I)	.803	.933	.964	.995	.803	.732	.720	.685	.803	.861	.868	.872
Ours: RadIR (I)→CXA (A)	.808	**.943**	**.985**	**1.000**	.808	**.782**	.751	.714	.808	**.870**	**.879**	**.882**
Negative Findings												
BiomedCLIP (I)	.716	.931	.983	.997	.716	.705	.689	.676	.716	.796	.823	.828
BiomedCLIP (T)	.716	.938	.965	.997	.716	.713	.691	.676	.716	.796	.821	.827
MedCLIP-ResNet (T)	.709	.952	.969	**1.000**	.709	.699	.675	.650	.734	.803	.828	.834
MedCLIP-ViT (T)	.692	.938	.976	.997	.692	.690	.669	.647	.709	.792	.818	.824
MedSigLIP (I)	.758	.945	.962	.993	.758	.753	.738	.711	.758	.824	.845	.850
MedSigLIP (T)	.730	.948	**.990**	**1.000**	.730	.716	.718	.707	.730	.822	.841	.845
XRaySigLIP (I)	.803	.962	.983	.997	.803	.765	.762	.739	.803	.862	.880	.882
XRaySigLIP (T)	.775	.948	**.990**	**1.000**	.775	.700	.687	.668	.775	.824	.853	.859
CheXficient (I)	.765	.952	.972	**1.000**	.765	.747	.743	.729	.765	.848	.853	.857
CheXficient (T)	.834	**.969**	.983	.990	.834	**.834**	.813	**.796**	.834	.897	.901	.902
CheXficient (A)	.830	.962	.979	.990	.830	.829	**.814**	.793	.830	.893	.897	.898
RadIR (I)	.862	.962	**.990**	.993	.862	.811	.801	.783	.862	.907	.913	.917
Ours: RadIR (I)→CXA (T)	**.869**	**.969**	.986	.997	**.869**	.818	.807	.791	**.869**	**.915**	**.920**	**.921**
Ours: RadIR (I)→CXA (I)	**.869**	.965	.986	.993	**.869**	.810	.799	.785	**.869**	**.915**	.919	.920
Ours: RadIR (I)→CXA (A)	.810	.959	.986	**1.000**	.810	.800	.796	.790	.810	.879	.886	.888

Table 8: Retrieval performance on **Consolidation** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.502	.540	.560	.620	.502	.495	.498	.499	.502	.521	.521	.534
BiomedCLIP (T)	.579	.700	.720	.800	.579	.567	.545	.542	.579	.609	.629	.645
MedCLIP-ResNet (T)	.583	.620	.700	.740	.583	.537	.538	.515	.561	.590	.597	.609
MedCLIP-ViT (T)	.574	.620	.680	.760	.574	.527	.526	.517	.579	.599	.611	.622
MedSigLIP (I)	.568	.639	.660	.800	.568	.528	.519	.527	.568	.583	.602	.621
MedSigLIP (T)	.577	.700	.720	.840	.577	.587	.548	.532	.577	.599	.625	.644
XRaySigLIP (I)	.567	.640	.660	.720	.567	.533	.522	.522	.567	.582	.596	.609
XRaySigLIP (T)	.593	.620	.660	.740	.593	.514	.507	.506	.593	.605	.611	.625
CheXficient (I)	.587	.639	.680	.740	.587	.537	.531	.518	.587	.610	.620	.627
CheXficient (T)	.561	.680	.720	.820	.561	.552	.540	.542	.561	.613	.622	.638
CheXficient (A)	.600	.680	.720	.760	.600	.558	.542	.527	.600	.637	.646	.651
RadIR (I)	.556	.640	.720	.860	.556	.525	.537	.548	.556	.586	.604	.629
Ours: RadIR (I)→CXA (T)	.560	.680	.700	.860	.560	.542	.535	.549	.560	.610	.614	.635
Ours: RadIR (I)→CXA (I)	.553	.640	.740	.840	.553	.524	.538	.540	.553	.593	.614	.628
Ours: RadIR (I)→CXA (A)	**.629**	**.839**	**.840**	**.900**	**.629**	**.609**	**.578**	**.553**	**.629**	**.721**	**.721**	**.727**
Positive Findings												
BiomedCLIP (I)	.040	.080	.120	.240	.040	.027	.024	.028	.040	.060	.060	.086
BiomedCLIP (T)	.200	.400	.440	.600	.200	.173	.136	.128	.200	.240	.280	.312
MedCLIP-ResNet (T)	.200	.240	.400	.480	.200	.120	.120	.076	.160	.200	.213	.237
MedCLIP-ViT (T)	.200	.240	.360	.520	.200	.107	.104	.084	.200	.220	.243	.265
MedSigLIP (I)	.160	.280	.320	.600	.160	.093	.072	.092	.160	.180	.217	.255
MedSigLIP (T)	.200	.400	.440	.680	.200	.213	.144	.112	.200	.220	.273	.312
XRaySigLIP (I)	.160	.280	.320	.440	.160	.093	.072	.072	.160	.180	.207	.231
XRaySigLIP (T)	.240	.240	.320	.480	.240	.080	.064	.060	.240	.240	.250	.279
CheXficient (I)	.200	.280	.360	.480	.200	.107	.096	.068	.200	.233	.253	.269
CheXficient (T)	.160	.360	.440	.640	.160	.147	.120	.124	.160	.247	.265	.296
CheXficient (A)	.240	.360	.440	.520	.240	.147	.120	.088	.240	.293	.311	.322
RadIR (I)	.160	.280	.440	.720	.160	.093	.120	.136	.160	.200	.233	.283
Ours: RadIR (I)→CXA (T)	.160	.360	.400	.720	.160	.120	.112	.136	.160	.240	.248	.290
Ours: RadIR (I)→CXA (I)	.160	.280	.480	.680	.160	.093	.120	.120	.160	.213	.257	.284
Ours: RadIR (I)→CXA (A)	**.280**	**.680**	**.680**	**.800**	**.280**	**.253**	**.184**	**.140**	**.280**	**.453**	**.453**	**.466**
Negative Findings												
BiomedCLIP (I)	.965	**1.000**	**1.000**	**1.000**	.965	.964	**.972**	.969	.965	.983	.983	.983
BiomedCLIP (T)	.958	**1.000**	**1.000**	**1.000**	.958	.961	.954	.957	.958	.977	.979	.979
MedCLIP-ResNet (T)	.965	**1.000**	**1.000**	**1.000**	.965	.954	.956	.953	.963	.979	.981	.981
MedCLIP-ViT (T)	.948	**1.000**	**1.000**	**1.000**	.948	.947	.949	.951	.958	.977	.979	.979
MedSigLIP (I)	.976	.998	**1.000**	**1.000**	.976	.962	.966	.962	.976	.987	.987	.987
MedSigLIP (T)	.954	**1.000**	**1.000**	**1.000**	.954	.960	.952	.953	.954	.977	.977	.977
XRaySigLIP (I)	.974	**1.000**	**1.000**	**1.000**	.974	**.973**	.971	**.973**	.974	.985	.986	.986
XRaySigLIP (T)	.945	**1.000**	**1.000**	**1.000**	.945	.949	.951	.952	.945	.971	.972	.972
CheXficient (I)	.974	.998	**1.000**	**1.000**	.974	.966	.965	.968	.974	.986	.986	.986
CheXficient (T)	.963	**1.000**	**1.000**	**1.000**	.963	.958	.960	.960	.963	.980	.980	.980
CheXficient (A)	.961	**1.000**	**1.000**	**1.000**	.961	.969	.965	.966	.961	.980	.980	.980
RadIR (I)	.952	**1.000**	**1.000**	**1.000**	.952	.957	.955	.959	.952	.973	.975	.975
Ours: RadIR (I)→CXA (T)	.961	**1.000**	**1.000**	**1.000**	.961	.964	.958	.962	.961	.979	.979	.979
Ours: RadIR (I)→CXA (I)	.945	**1.000**	**1.000**	**1.000**	.945	.955	.956	.961	.945	.972	.972	.972
Ours: RadIR (I)→CXA (A)	**.978**	.998	**1.000**	**1.000**	**.978**	.964	.971	.966	**.978**	**.988**	**.988**	**.988**

Table 9: Retrieval performance on **Edema** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.668	.855	.933	**.995**	.668	.620	.609	.601	.668	.744	.768	.781
BiomedCLIP (T)	.714	.886	.939	.970	.714	.687	.666	.639	.714	.781	.802	.810
MedCLIP-ResNet (T)	.625	.836	.874	.944	.625	.608	.587	.572	.620	.684	.716	.730
MedCLIP-ViT (T)	.603	.802	.864	.950	.603	.591	.591	.578	.615	.686	.719	.732
MedSigLIP (I)	.771	.906	.946	.970	.771	.710	.679	.662	.771	.828	.842	.848
MedSigLIP (T)	.726	.828	.879	.960	.726	.682	.671	.658	.726	.767	.780	.796
XRaySigLIP (I)	.771	.897	.958	.975	.771	.690	.663	.641	.771	.811	.833	.843
XRaySigLIP (T)	.622	.828	.913	.970	.622	.593	.592	.563	.622	.700	.732	.744
CheXficient (I)	.771	.912	.952	.980	.771	.708	.684	.657	.771	.832	.841	.845
CheXficient (T)	.762	.932	.948	.980	.762	.743	.716	.685	.762	.838	.842	.846
CheXficient (A)	.800	.933	.962	.980	.800	.758	.723	.694	.800	.861	.868	.871
RadIR (I)	.753	.887	.968	.990	.753	.670	.660	.641	.753	.792	.818	.831
Ours: RadIR (I)→CXA (T)	.747	.883	.958	.985	.747	.678	.655	.646	.747	.807	.824	.828
Ours: RadIR (I)→CXA (I)	.743	.906	.968	.990	.743	.680	.652	.648	.743	.815	.830	.833
Ours: RadIR (I)→CXA (A)	**.862**	**.952**	**.981**	**.995**	**.862**	**.804**	**.794**	**.752**	**.862**	**.901**	**.907**	**.909**
Positive Findings												
BiomedCLIP (I)	.465	.727	.869	**.990**	.465	.404	.378	.370	.465	.571	.610	.636
BiomedCLIP (T)	.556	.808	.879	.939	.556	.525	.487	.449	.556	.657	.683	.698
MedCLIP-ResNet (T)	.404	.677	.748	.889	.404	.384	.339	.313	.394	.460	.514	.543
MedCLIP-ViT (T)	.343	.606	.727	.899	.343	.337	.331	.308	.394	.475	.528	.554
MedSigLIP (I)	.657	.838	.899	.939	.657	.559	.507	.481	.657	.737	.757	.764
MedSigLIP (T)	.556	.677	.768	.919	.556	.478	.475	.455	.556	.601	.621	.653
XRaySigLIP (I)	.657	.818	.919	.950	.657	.525	.489	.462	.657	.707	.737	.754
XRaySigLIP (T)	.414	.677	.828	.939	.414	.377	.370	.321	.414	.505	.560	.582
CheXficient (I)	.657	.859	.909	.960	.657	.545	.515	.473	.657	.741	.753	.759
CheXficient (T)	.626	.879	.899	.960	.626	.596	.564	.502	.626	.741	.746	.754
CheXficient (A)	.697	.879	.929	.960	.697	.626	.572	.525	.697	.781	.793	.797
RadIR (I)	.616	.798	.939	.980	.616	.481	.469	.450	.616	.662	.702	.728
Ours: RadIR (I)→CXA (T)	.606	.788	.919	.970	.606	.498	.465	.459	.606	.683	.713	.721
Ours: RadIR (I)→CXA (I)	.606	.838	.939	.980	.606	.505	.461	.458	.606	.707	.730	.736
Ours: RadIR (I)→CXA (A)	**.818**	**.919**	**.970**	**.990**	**.818**	**.721**	**.711**	**.656**	**.818**	**.860**	**.871**	**.874**
Negative Findings												
BiomedCLIP (I)	.872	.982	.997	**1.000**	.872	.836	.839	.832	.872	.916	.925	.927
BiomedCLIP (T)	.872	.963	**1.000**	**1.000**	.872	.849	.845	.830	.872	.905	.920	.922
MedCLIP-ResNet (T)	.846	.995	**1.000**	**1.000**	.846	.832	.834	.831	.846	.909	.918	.918
MedCLIP-ViT (T)	.862	**.997**	**1.000**	**1.000**	.862	.845	.851	.847	.836	.898	.909	.910
MedSigLIP (I)	.885	.974	.992	**1.000**	.885	.861	.850	.842	.885	.918	.928	.931
MedSigLIP (T)	.896	.979	.990	**1.000**	.896	.885	.868	.861	.896	.933	.938	.940
XRaySigLIP (I)	.885	.977	.997	**1.000**	.885	.855	.837	.820	.885	.915	.929	.931
XRaySigLIP (T)	.830	.979	.997	**1.000**	.830	.808	.814	.805	.830	.894	.903	.906
CheXficient (I)	.885	.966	.995	**1.000**	.885	.870	.853	.841	.885	.922	.929	.930
CheXficient (T)	.898	.984	.997	**1.000**	.898	**.889**	.869	**.869**	.898	.936	.939	.939
CheXficient (A)	.903	.987	.995	**1.000**	.903	**.889**	.874	.863	.903	.941	.943	.944
RadIR (I)	.890	.977	.997	**1.000**	.890	.859	.852	.833	.890	.922	.934	.935
Ours: RadIR (I)→CXA (T)	.888	.979	.997	**1.000**	.888	.857	.845	.833	.888	.930	.934	.935
Ours: RadIR (I)→CXA (I)	.880	.974	.997	**1.000**	.880	.856	.844	.839	.880	.923	.929	.929
Ours: RadIR (I)→CXA (A)	**.906**	.984	.992	**1.000**	**.906**	.887	**.878**	.849	**.906**	**.942**	**.944**	**.945**

Table 10: Retrieval performance on **Cardiomegaly** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.612	.858	.916	.977	.612	.604	.586	.578	.612	.702	.735	.746
BiomedCLIP (T)	.719	.894	.930	.986	.719	.635	.605	.585	.719	.783	.805	.815
MedCLIP-ResNet (T)	.606	.836	.900	.977	.606	.600	.586	.568	.632	.684	.718	.740
MedCLIP-ViT (T)	.598	.804	.888	.972	.598	.583	.571	.575	.622	.686	.722	.736
MedSigLIP (I)	.642	.876	.932	.977	.642	.620	.606	.606	.642	.726	.756	.767
MedSigLIP (T)	.634	.789	.869	.958	.634	.605	.599	.584	.634	.689	.717	.734
XRaySigLIP (I)	.657	.891	.943	.986	.657	.620	.609	.592	.657	.732	.769	.778
XRaySigLIP (T)	.624	.827	.888	.958	.624	.593	.573	.543	.624	.690	.723	.736
CheXficient (I)	.648	.892	.935	.981	.648	.631	.618	.605	.648	.756	.766	.772
CheXficient (T)	.754	.929	**.975**	.981	.754	.719	**.705**	**.682**	.754	.836	.847	.848
CheXficient (A)	.714	.908	.962	.986	.714	.680	.685	.666	.714	.800	.813	.817
RadIR (I)	.776	.933	.962	.985	.776	.704	.683	.666	.776	.816	.845	.851
Ours: RadIR (I)→CXA (T)	.767	.931	.967	.975	.767	.700	.673	.657	.767	.837	.844	.846
Ours: RadIR (I)→CXA (I)	.786	.931	.959	.985	.786	.694	.686	.665	.786	.847	.854	.857
Ours: RadIR (I)→CXA (A)	**.801**	**.947**	.974	**.995**	**.801**	**.735**	.703	.669	**.801**	**.867**	**.873**	**.876**
Positive Findings												
BiomedCLIP (I)	.402	.748	.832	.953	.402	.405	.372	.360	.402	.523	.574	.594
BiomedCLIP (T)	.598	.813	.860	.972	.598	.467	.407	.381	.598	.678	.703	.722
MedCLIP-ResNet (T)	.393	.701	.804	.953	.393	.389	.364	.324	.430	.486	.534	.577
MedCLIP-ViT (T)	.364	.626	.785	.944	.364	.346	.336	.345	.402	.477	.533	.561
MedSigLIP (I)	.430	.776	.869	.953	.430	.405	.385	.386	.430	.556	.601	.622
MedSigLIP (T)	.393	.589	.738	.916	.393	.346	.344	.332	.393	.453	.502	.536
XRaySigLIP (I)	.458	.804	.888	.972	.458	.430	.407	.379	.458	.565	.623	.640
XRaySigLIP (T)	.393	.673	.776	.916	.393	.352	.320	.279	.393	.477	.528	.554
CheXficient (I)	.458	.804	.878	.963	.458	.436	.406	.378	.458	.611	.627	.639
CheXficient (T)	.626	.878	**.953**	.963	.626	.573	**.553**	**.520**	.626	.745	.763	.764
CheXficient (A)	.561	.832	.935	.972	.561	.505	.512	.482	.561	.681	.704	.710
RadIR (I)	.673	.888	.935	.972	.673	.551	.510	.484	.673	.715	.763	.774
Ours: RadIR (I)→CXA (T)	.654	.878	.944	.953	.654	.542	.495	.465	.654	.746	.760	.762
Ours: RadIR (I)→CXA (I)	.692	.878	.925	.972	.692	.530	.518	.482	.692	.768	.779	.786
Ours: RadIR (I)→CXA (A)	**.701**	**.916**	**.953**	**.991**	**.701**	**.604**	.544	.476	**.701**	**.797**	**.806**	**.812**
Negative Findings												
BiomedCLIP (I)	.821	.968	**1.000**	**1.000**	.821	.803	.800	.796	.821	.880	.896	.897
BiomedCLIP (T)	.840	.976	**1.000**	**1.000**	.840	.803	.803	.789	.840	.888	.907	.907
MedCLIP-ResNet (T)	.819	.971	.997	**1.000**	.819	.811	.807	.812	.835	.881	.902	.903
MedCLIP-ViT (T)	.832	.981	.992	**1.000**	.832	.820	.806	.804	.843	.896	.911	.912
MedSigLIP (I)	.853	.976	.995	**1.000**	.853	.835	.828	.825	.853	.896	.911	.913
MedSigLIP (T)	.875	**.989**	**1.000**	**1.000**	.875	.863	.854	.836	.875	.924	.932	.932
XRaySigLIP (I)	.856	.979	.997	**1.000**	.856	.810	.811	.805	.856	.899	.915	.916
XRaySigLIP (T)	.856	.981	**1.000**	**1.000**	.856	.833	.826	.806	.856	.904	.917	.918
CheXficient (I)	.837	.981	.992	**1.000**	.837	.826	.831	.832	.837	.902	.905	.906
CheXficient (T)	.883	.979	.997	**1.000**	.883	**.865**	.856	.845	.883	.927	.931	.931
CheXficient (A)	.867	.984	.989	**1.000**	.867	.856	.858	.850	.867	.920	.922	.923
RadIR (I)	.880	.979	.989	.997	.880	.856	.856	.849	.880	.917	.927	.929
Ours: RadIR (I)→CXA (T)	.880	.984	.989	.997	.880	.857	.851	.850	.880	.927	.928	.929
Ours: RadIR (I)→CXA (I)	.880	.984	.992	.997	.880	.859	.854	.847	.880	.926	.928	.929
Ours: RadIR (I)→CXA (A)	**.901**	.979	.995	**1.000**	**.901**	**.865**	**.861**	**.862**	**.901**	**.936**	**.940**	**.941**

Table 11: Retrieval performance on **Atelectasis** across all $k \in \{1, 3, 5, 10\}$. Recall@k denotes hit rate; Prec@k denotes precision; MRR@k denotes mean reciprocal rank. *Macro* averages positive- and negative-class results; *Positive* and *Negative* report retrieval over the corresponding class subset. Stage 2 (*Ours*) reranks RadIR (I) candidates with token-level optimal transport. Best Stage-1 score per column is underlined; best overall is **bold**. CXA=CheXagent; (I)=image+text input, unfiltered; (T)=text-only input; (A)=image+text input, filtered to visually-enriched text tokens.

	Recall@k				Prec@k				MRR@k			
Method	1	3	5	10	1	3	5	10	1	3	5	10
Macro (A)veraged												
BiomedCLIP (I)	.615	.869	.950	**.996**	.615	.585	.561	.542	.615	.690	.739	.750
BiomedCLIP (T)	.651	.880	.958	**.996**	.651	.604	.587	.574	.651	.715	.762	.772
MedCLIP-ResNet (T)	.637	.858	.945	.989	.637	.592	.590	.582	.653	.721	.757	.768
MedCLIP-ViT (T)	.645	.852	.915	.982	.645	.593	.586	.570	.641	.710	.747	.762
MedSigLIP (I)	.624	.873	.948	.986	.624	.555	.559	.547	.624	.707	.744	.757
MedSigLIP (T)	.645	.863	.925	.971	.645	.588	.565	.549	.645	.716	.751	.763
XRaySigLIP (I)	.625	.861	.942	.986	.625	.569	.555	.543	.625	.711	.747	.757
XRaySigLIP (T)	.636	.827	.907	.964	.636	.587	.571	.539	.635	.704	.732	.748
CheXficient (I)	.646	.875	.935	.982	.646	.575	.555	.546	.646	.746	.760	.766
CheXficient (T)	.671	.891	.952	.993	.671	.618	.599	.595	.671	.770	.784	.789
CheXficient (A)	.644	.880	.941	.982	.644	.614	.593	.585	.644	.751	.766	.771
RadIR (I)	.661	.882	.941	.989	.661	.606	.600	.579	.661	.729	.765	.778
Ours: RadIR (I)→CXA (T)	.693	.870	.941	.986	.693	.617	.606	.584	.693	.771	.787	.793
Ours: RadIR (I)→CXA (I)	.647	.874	.942	.986	.647	.606	.595	.580	.647	.747	.763	.769
Ours: RadIR (I)→CXA (A)	**.737**	**.902**	**.960**	**.996**	**.737**	**.687**	**.648**	**.607**	**.737**	**.811**	**.824**	**.829**
Positive Findings												
BiomedCLIP (I)	.475	.755	.899	**.993**	.475	.429	.403	.381	.475	.547	.618	.640
BiomedCLIP (T)	.532	.799	.921	**.993**	.532	.477	.465	**.450**	.532	.590	.662	.678
MedCLIP-ResNet (T)	.489	.748	.899	.978	.489	.432	.437	.420	.511	.586	.638	.656
MedCLIP-ViT (T)	.496	.741	.842	.964	.496	.441	.434	.409	.504	.583	.631	.657
MedSigLIP (I)	.446	.763	.899	.971	.446	.355	.370	.355	.446	.550	.603	.627
MedSigLIP (T)	.496	.755	.856	.942	.496	.410	.371	.356	.496	.572	.622	.644
XRaySigLIP (I)	.475	.748	.885	.971	.475	.384	.357	.345	.475	.572	.622	.642
XRaySigLIP (T)	.446	.676	.820	.928	.446	.377	.352	.322	.446	.525	.565	.596
CheXficient (I)	.475	.770	.871	.964	.475	.393	.363	.343	.475	.601	.625	.637
CheXficient (T)	.511	.806	.906	.986	.511	.448	.413	.418	.511	.643	.666	.677
CheXficient (A)	.460	.784	.885	.964	.460	.424	.397	.386	.460	.608	.632	.643
RadIR (I)	.504	.770	.885	.978	.504	.436	.424	.391	.504	.590	.635	.660
Ours: RadIR (I)→CXA (T)	.532	.748	.885	.971	.532	.436	.422	.391	.532	.627	.658	.670
Ours: RadIR (I)→CXA (I)	.475	.763	.885	.971	.475	.432	.419	.394	.475	.601	.629	.641
Ours: RadIR (I)→CXA (A)	**.597**	**.827**	**.928**	**.993**	**.597**	**.544**	**.499**	.413	**.597**	**.699**	**.722**	**.730**
Negative Findings												
BiomedCLIP (I)	.755	.983	**1.000**	**1.000**	.755	.742	.720	.703	.755	.832	.861	.861
BiomedCLIP (T)	.770	.962	.994	**1.000**	.770	.731	.710	.699	.770	.840	.862	.865
MedCLIP-ResNet (T)	.784	.968	.991	**1.000**	.784	.753	.743	.743	.796	.856	.876	.879
MedCLIP-ViT (T)	.793	.962	.988	**1.000**	.793	.745	.738	.732	.778	.838	.863	.867
MedSigLIP (I)	.802	.983	.997	**1.000**	.802	.755	.749	.740	.802	.864	.885	.887
MedSigLIP (T)	.793	.971	.994	**1.000**	.793	.767	.758	.741	.793	.860	.879	.881
XRaySigLIP (I)	.775	.974	**1.000**	**1.000**	.775	.755	.753	.741	.775	.850	.872	.873
XRaySigLIP (T)	.825	.977	.994	**1.000**	.825	.798	.789	.757	.825	.883	.899	.900
CheXficient (I)	.816	.980	**1.000**	**1.000**	.816	.756	.747	.749	.816	.891	.896	.896
CheXficient (T)	.831	.977	.997	**1.000**	.831	.787	.784	.771	.831	.896	.901	.902
CheXficient (A)	.828	.977	.997	**1.000**	.828	.803	.789	.784	.828	.894	.899	.900
RadIR (I)	.819	**.994**	.997	**1.000**	.819	.775	.775	.766	.819	.869	.895	.895
Ours: RadIR (I)→CXA (T)	.854	.991	.997	**1.000**	.854	.798	.789	.778	.854	.914	.915	.916
Ours: RadIR (I)→CXA (I)	.819	.985	**1.000**	**1.000**	.819	.780	.771	.766	.819	.893	.897	.897
Ours: RadIR (I)→CXA (A)	**.878**	.977	.991	**1.000**	**.878**	**.829**	**.797**	**.800**	**.878**	**.923**	**.926**	**.927**