

RadIR: A Scalable Framework for Multi-Grained Medical Image Retrieval via Radiology Report Mining

Tengfei Zhang^{1,2*}, Ziheng Zhao^{2,3*}, Chaoyi Wu^{2,3}, Xiao Zhou²,
Ya Zhang^{2,3}, Yanfeng Wang^{2,3†}, and Weidi Xie^{2,3†}

¹University of Science and Technology of China

²Shanghai AI Laboratory

³Shanghai Jiao Tong University

{wangyanfeng622, weidi}@sjtu.edu.cn

Abstract. Developing advanced medical imaging retrieval systems is challenging due to the varying definitions of ‘similar images’ across different medical contexts. This challenge is compounded by the lack of large-scale, high-quality medical imaging retrieval datasets and benchmarks. In this paper, we propose a novel methodology that leverages dense radiology reports to define image-wise similarity ordering at multiple granularities in a scalable and fully automatic manner. Using this approach, we construct two comprehensive medical imaging retrieval datasets: **MIMIC-IR** for Chest X-rays and **CTRATE-IR** for CT scans, providing detailed image-image ranking annotations conditioned on diverse anatomical structures. Furthermore, we develop two retrieval systems, **RadIR-CXR** and **RadIR-ChestCT**, which demonstrate superior performance in traditional image-image and image-report retrieval tasks. These systems also enable flexible, effective image retrieval conditioned on specific anatomical structures described in text, achieving state-of-the-art results on 77 out of 78 metrics.

Keywords: Image Retrieval · Medical Imaging · Vision-Language Pre-training.

1 Introduction

The objective of this paper is to develop an image retrieval system for medical applications that ranks instances in a retrieval set based on their relevance to a query, which includes a radiology image and an optional text condition indicating the region to focus on, *i.e.*, the name of anatomy. Such a system has broad implications in enhancing clinicians’ ability to identify similar cases, supporting diagnosis and treatment planning, and facilitating medical education and research [2,17,15,5]. Furthermore, in building generalist models [7,20,22],

* Equal Contribution. † Corresponding Author

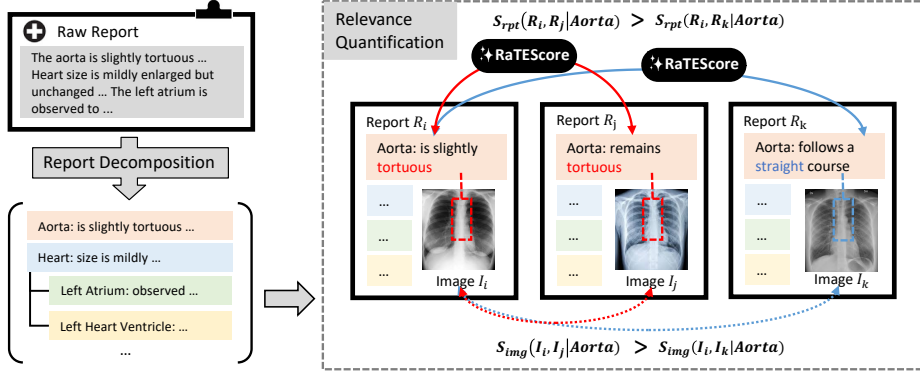


Fig. 1. Fine-grained image similarity derived from report. We decompose report into anatomy-centric findings and leveraging state-of-the-art medical language model RaTEScore to assess their relevance. We treat this as a proxy for the fine-grained image-image similarity, preserving their rankings in clinical meanings.

retrieval-augmented generation (RAG) plays a crucial role in reducing hallucinations and supporting case-based reasoning by grounding outputs in retrieved evidence.

Developing medical image retrieval systems is particularly challenging due to the complexity of defining image similarity, which depends on multiple factors such as global appearance, localized findings, and specific pathologies. For instance, two patients with different diseases may exhibit similar localized abnormalities. Capturing these nuanced relationships requires a granular understanding beyond coarse pathological or image-level labels. However, manual annotation of fine-grained similarity is often impractical due to its labor-intensive and subjective nature, especially at the scale needed for large datasets. Existing benchmarks [1,4,9,12,13] typically rely on coarse image-level labels or limited manual annotations, which fail to capture the full spectrum of clinically relevant features, thereby limiting the development of scalable systems.

To address the challenges we propose a novel medical image ranking pipeline by mining the multi-grained annotations from corresponding radiology reports. Specifically, given a certain anatomy structure, we first standardize the paired reports and extract the relative findings. Then, we adopt the text-level similarity ranking of the findings based on well-designed language-wise metrics [23], to, in turn, represent the image similarity ranks regarding this anatomy structure. This pipeline enables the construction of multi-granular similarity ranking training data in a scalable and automated manner, for both global image matching and fine-grained retrieval conditioned on anatomy structures, as shown in Figure 1. Based on it, we extend two widely used datasets, MIMIC-CXR [11] and CT-RATE [6], to create two large-scale image retrieval datasets, **MIMIC-IR** and **CTRATE-IR**, with detailed annotations of image-image similarity ordering mined from dense report annotation, serving for both training and evaluation.

On model development, leveraging the two datasets, we have trained two retrieval systems: **RadIR-CXR** and **RadIR-ChestCT**. These systems achieve state-of-the-art performance in traditional image and image-report retrieval, while further enabling fine-grained retrieval with anatomy terminology as text condition. They allow users to query specific anatomies, bridging the gap between global similarity and localized retrieval, thus better fitting clinical demands.

In summary, our contributions are threefold: (i) We propose a novel, fully automated pipeline to structure radiology reports and bridge multi-grained image-image relevance in a scalable manner. (ii) We develop **MIMIC-IR** and **CTRATE-IR**, two large-scale and comprehensive datasets accompanied by evaluation benchmarks for Chest X-ray and Chest CT image retrieval, with detailed annotations capturing image-image similarity ordering based on regional findings. (iii) We present two state-of-the-art image retrieval systems, **RadIR-CXR** and **RadIR-ChestCT**, which demonstrate superior performance in global image retrieval and substantial advancements in image retrieval conditioned on anatomies.

2 Problem Formulation

Considering a collection of radiology image-report pairs, denoted as $\mathcal{D} = \{(I_1, R_1), \dots, (I_K, R_K)\}$, where $I_i \in \mathbb{R}^{H \times W \times C}$ refers to the radiology image, and R_i is the corresponding clinical report. The goal of the image retrieval task is to find the similar cases from \mathcal{D} , given a query image I_q and optionally, a conditional query Q referring to an anatomical structure. This is equivalent to ranking the candidates in \mathcal{D} based on their relevance to the query image:

$$\{r_1, r_2, \dots, r_K\} = \mathcal{I}(\mathcal{S}_{\text{img}}(I_q, I_j \mid Q)), \forall I_j \in \mathcal{D} \quad (1)$$

where r_i denotes the rankings, and $\mathcal{I}(\cdot)$ is a function that indexes the image similarity $\mathcal{S}_{\text{img}}(\cdot)$. When Q is not provided, this reduces to a conventional image retrieval task without any conditions.

Discussion. In this ranking task, estimating the exact similarity values between images is unnecessary. Instead, we focus on preserving the relative similarity ordering. In this paper, we make the assumption that radiology reports have faithfully captured the critical findings of their paired images. Consequently, the similarity ranking of images should align with the similarity ranking of their corresponding reports. Thus, we use the similarity between radiology reports, denoted as $\mathcal{F}_{\text{rpt}}(\cdot)$, as a feasible and practical proxy for image similarity:

$$\mathcal{I}(\mathcal{S}_{\text{img}}(I_q, I_j \mid Q)) = \mathcal{I}(\mathcal{S}_{\text{rpt}}(R_q, R_j \mid Q)), \forall I_j, R_j \in \mathcal{D} \quad (2)$$

The following sections detail the procedure for quantifying similarity between reports and leveraging these rankings to train the image retrieval system.

3 Dataset Construction

In this section, we propose an automatic pipeline to quantify image-to-image similarity ordering, via mining their paired radiology reports, as shown in Figure 1. We first introduce the data sources in Section 3.1. Then, we detail the

two main procedures in the pipeline: report decomposition in Section 3.2 and relevance quantification in Section 3.3.

3.1 Data Sources

We utilize two widely used datasets: **MIMIC-CXR** [11] is the largest chest X-ray dataset, containing **377,110** image-report pairs; while **CT-RATE** [6] is a large-scale chest CT dataset with **25,692** non-contrast CT volume-report pairs. These reports include detailed radiological findings and impressions, which are essential for defining clinically meaningful similarities between images.

3.2 Report Decomposition

We describe our process for extracting and structuring anatomical regions and their associated findings from radiology reports. This involves building a comprehensive anatomy terminology set, extracting regional findings, and integrating hierarchical relationships between anatomical structures, as detailed below.

Anatomy Terminology Set. We utilized RadGraph-XL [3] to extract anatomical structures from radiology reports. A total of 90 high-frequency anatomical structures commonly referenced in radiology were identified. To ensure consistency, synonymous terms (*e.g.*, “superior vena cava” and “SVC”) were unified. The anatomical structures were further organized into a hierarchical framework, capturing relationships between parent structures (*e.g.*, “lungs”) and their substructures (*e.g.*, “left lung” and “right lung”).

Regional Findings Extraction. From the ‘Findings’ section of the reports, we extract region-specific findings by segmenting the content into sentences with periods as delimiters, and linking each sentence to the anatomical structures it mentions based on the anatomy terminology set.

Hierarchical Structure Integration. Relationships between anatomies, such as “lungs” and “left lung”, are utilized to merge findings from substructures into their parent structures. This integration provides a comprehensive, multi-level representation of findings for each anatomical region.

3.3 Relevance Quantification

After performing fine-grained report decomposition, we can further quantify the relevance between findings from different reports regarding the same anatomy, as a substitute for the corresponding fine-grained image similarity on it. Here, we apply RaTEScore [23], a state-of-the-art model that provides a robust evaluation metric for radiology report texts similarity based on key entities, as a proxy:

$$\mathcal{S}_{\text{rpt}}(R_q, R_r \mid Q) = \text{RaTEScore}(\mathcal{E}(R_q \mid Q), \mathcal{E}(R_r \mid Q)) \quad (3)$$

where \mathcal{E} denotes extracting regional findings from the raw reports regarding Q heuristically with rule-based string matching. For image retrieval without specific

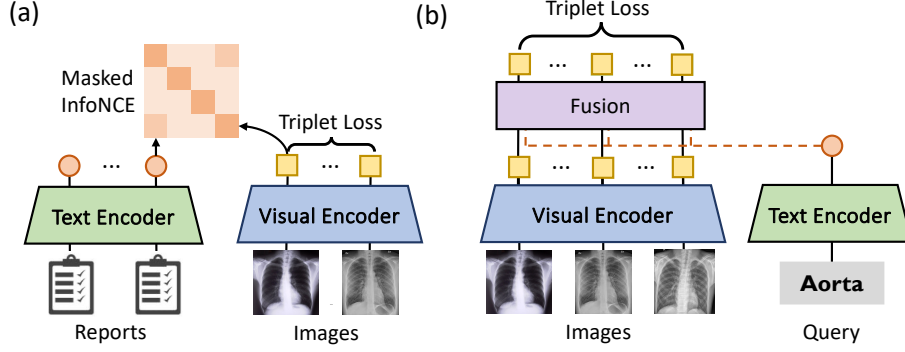


Fig. 2. Architecture and training procedures of RadIR. (a) In stage 1, we pre-train a CLIP-style model for unconditional image and image-report retrieval; (b) In stage 2, we extend the pre-trained model for image retrieval conditioned on anatomies.

query conditions that Q is empty, \mathcal{E} will return the original report that evaluates the similarity of entire reports as a substitute for the global image similarity.

Summary. We extract **2,582,477** regional findings in total, covering **90** anatomical structures. We further quantify over **132 billion** fine-grained image-image relevance between them. We name the two proposed large-scale and multi-granularity datasets as **MIMIC-IR** and **CTRATE-IR**, as the foundation to train and benchmark the radiology image retrieval systems.

4 RadIR

In this section, we present the details to build **RadIR** based on the datasets we construct above. The training procedure includes two stages. In Section 4.1, we pre-train the CLIP-style model for unconditional image retrieval; In Section 4.2, we extend the pre-trained model for retrieval task conditioned on a text query.

4.1 Unconditional Image Retrieval

Architecture. As shown in Figure 2(a), in this setting, we directly encode the raw images and reports without considering extra text queries. We adopt a typical CLIP-style [18] model with a Vision Transformer based image encoder $\Phi_{\text{visual}}(\cdot)$ and a BERT-based text encoder $\Phi_{\text{text}}(\cdot)$:

$$v = \Phi_{\text{visual}}(I) \in \mathbb{R}^d, \quad t = \Phi_{\text{text}}(R) \in \mathbb{R}^d \quad (4)$$

where I denotes a radiology image, R denotes a radiology report, v and t denotes their features respectively, and d is the dimension.

Table 1. Unconditional image to image, and image to report retrieval results. Recall and NDCG results are presented in percentage. The best results on each metric are bolded.

Method	<i>Recall@k</i> ↑				<i>NDCG</i> ↑			
	k=5	k=10	k=50	k=100	k=5	k=10	k=50	k=100
<i>on MIMIC-IR (Chest X-Ray)</i>								
Image2Image								
MedCLIP	3.05	4.77	12.65	18.93	67.15	44.49	16.74	10.70
BioMedCLIP	2.04	3.30	8.20	12.68	64.49	42.72	16.07	10.27
PMC-CLIP	2.20	3.58	8.07	12.03	63.23	41.88	15.75	10.06
RadIR-CXR	5.18	6.94	15.45	21.29	68.23	45.21	17.01	10.88
Image2Text								
MedCLIP	0.19	0.28	2.04	3.77	58.22	38.58	14.51	9.27
BioMedCLIP	0.47	0.78	4.23	8.10	62.79	41.60	15.64	9.99
PMC-CLIP	0.31	0.44	2.73	5.43	50.96	35.18	13.99	9.06
RadIR-CXR	4.33	6.88	18.18	25.34	69.07	45.76	17.21	11.00
<i>on CTRATE-IR (Chest CT)</i>								
Image2Image								
CT-CLIP	19.43	28.76	57.51	68.13	74.48	75.20	78.05	79.96
RadIR-ChestCT	20.75	30.57	62.44	72.80	74.60	75.47	78.51	80.49
Image2Text								
CT-CLIP	5.05	8.19	25.27	39.92	67.57	70.50	76.67	79.45
RadIR-ChestCT	6.65	12.99	36.72	52.91	69.18	72.11	78.12	80.84

Training Objectives. Given a batch of N samples, we can calculate the following similarity matrix as prediction:

$$\mathbf{S}_{i2t} = \mathbf{v}\mathbf{t}^T, \mathbf{S}_{t2i} = \mathbf{t}\mathbf{v}^T, \mathbf{S}_{i2i} = \mathbf{v}\mathbf{v}^T, \quad \mathbf{v} \mathbf{t} \in \mathbb{R}^{N \times d} \quad (5)$$

where \mathbf{S} denotes the similarity matrices from image-text and image-image, respectively, and \mathbf{v}, \mathbf{t} denotes the visual or text embedding set. Then, We applied masked infoNCE loss (MIL) [16] and triplet loss (TL) [8] to optimize our model:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MIL}}(\mathbf{S}_{i2t}, \mathbf{T}) + \lambda_2 \mathcal{L}_{\text{MIL}}(\mathbf{S}_{t2i}, \mathbf{T}) + \lambda_3 \mathcal{L}_{\text{TL}}(\mathbf{S}_{i2i}, \mathbf{T}) \quad (6)$$

where $\mathbf{T} \in \mathbb{R}^{N \times N}$ is a text-text similarity matrix calculated via RaTEScore, as illustrated in Section 3. $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters. \mathcal{L}_{MIL} is a variant of the classic infoNCE loss:

$$\mathcal{L}_{\text{MIL}}(\mathbf{S}, \mathbf{T}) = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\exp(\mathbf{S}_{ii})}{\sum_{j=1}^N \exp(\mathbf{S}_{ij}) \cdot (\mathbf{I} + \mathbf{1}[\mathbf{T} < \tau])_{ij}} \right) \quad (7)$$

where $\mathbf{1}[\mathbf{T} < \tau]$ is a matrix that masks out the potential positive elements outside the diagonal, based on a predefined threshold τ .

4.2 Text-Conditioned Image Retrieval

Architecture. As shown in Figure 2(b), given a conditional query Q , we employ a fusion module Φ_{fusion} to extend the model for conditional image retrieval:

$$f = \Phi_{\text{fusion}}(\Phi_{\text{visual}}(I), \Phi_{\text{text}}(Q)) \in \mathbb{R}^d \quad (8)$$

Table 2. Conditional image retrieval results on MIMIC-IR. Recall scores are averaged and aggregated by anatomical region, and presented in percentage. Anatomies are sorted in descending order of their frequency in train set, with the ‘head’ regions at the top and the ‘tail’ regions at the bottom. The best results for each anatomy are bolded. Greener suggests higher improvement over baselines.

Anatomy	Recall@3 \uparrow				Recall@5 \uparrow				Recall@10 \uparrow			
	PMC CLIP	BioMed CLIP	Med CLIP	RadIR	PMC CLIP	BioMed CLIP	Med CLIP	RadIR	PMC CLIP	BioMed CLIP	Med CLIP	RadIR
Pleura	16.11	18.67	21.98	25.32	23.51	28.35	30.44	32.99	34.09	40.57	40.52	44.60
Bones	12.06	17.24	13.43	18.79	20.97	24.74	21.91	26.16	34.85	35.04	35.23	38.09
Lung	7.08	9.48	8.62	11.37	11.55	12.54	14.88	16.46	18.03	19.03	22.41	22.31
Diaphragm	16.11	17.75	18.95	21.13	23.51	24.29	23.18	27.13	34.09	36.65	34.95	37.77
Vascular	14.80	19.57	22.93	30.65	29.04	27.97	32.63	36.39	39.96	39.85	45.76	47.37
Thorax	4.24	11.84	8.61	14.52	12.96	17.02	15.96	19.59	20.83	25.42	27.37	29.27
Heart	11.51	10.44	7.90	16.02	15.84	16.23	12.90	23.40	26.15	25.25	25.76	32.79
Airway	15.93	12.28	14.85	22.50	24.47	16.91	22.71	29.89	35.10	26.43	34.39	42.97
Stomach	7.41	11.85	12.59	19.23	10.37	15.56	20.00	22.22	22.96	23.70	24.44	31.11
Bronchi	16.83	13.86	13.86	28.71	20.79	16.83	20.79	31.68	32.67	22.77	38.61	44.55
Average	12.58	14.30	14.37	20.83	19.67	20.04	21.54	26.59	30.46	29.47	32.94	37.08

where f denotes the fused feature. This enables the model to capture relevant visual features based on the anatomy.

Training Objectives. In contrast to the global similarity matrix \mathbf{T} based on complete reports in Section 4.1, we introduce \mathbf{T}_Q as an anatomy-conditioned similarity matrix constructed from regional findings, based on equation 3. Meanwhile, the predicted conditional image-image similarity result \mathbf{S}_{f2f} is derived from the dot product of fused features. We then apply triplet loss on them:

$$\mathcal{L} = \mathcal{L}_{\text{TL}}(\mathbf{S}_{f2f}, \mathbf{T}_Q) \quad (9)$$

5 Experiment Settings and Results

We validate RadIR on MIMIC-IR and CTRATE-IR, with both unconditional image retrieval, image to report retrieval, and image retrieval conditioned on anatomy name. In all experiments, we follow the official train-test split of MIMIC-CXR and CT-RATE. In this section, we first introduce our baselines in Section 5.1 and benchmark metrics in Section 5.2; Then, we analyze the experiment results in Section 5.3 and Section 5.4.

5.1 Baseline

We take the following methods as baselines: **BioMedCLIP** [21], a vision-language foundation model for 2D biomedical images pre-trained on 15M image-text pairs; **MedCLIP** [19], a decoupled image-text contrastive learning framework for chest X-Ray images trained on MIMIC-CXR [11] and CheXpert [10]; **PMC-CLIP** [14], a CLIP-style model pretrained on PMC-OA with 1.6M biomedical

Table 3. Conditional image retrieval results on CTRATE-IR. Recall scores are in percentage. Anatomies are in descending order of their frequency in train set, with the ‘head’ anatomies at the top and the ‘tail’ anatomies at the bottom. The best results for each anatomy are bolded. Greener suggests higher improvement over baseline.

Anatomy	#Samples	Recall@3 \uparrow		Recall@5 \uparrow		Recall@10 \uparrow	
		CT-CLIP	RadIR	CT-CLIP	RadIR	CT-CLIP	RadIR
Bone	23.5k	45.75	49.76	56.33	60.51	67.31	71.03
Heart	23.3k	33.75	34.15	43.19	43.68	55.72	59.44
Bronchie	21.7k	55.18	57.76	67.20	69.42	75.81	78.43
Trachea	21.7k	57.43	60.48	69.24	70.51	77.71	80.61
Pleura	18.2k	35.14	40.57	44.59	54.14	60.00	71.64
Vertebrae	13.5k	57.69	62.04	63.69	66.91	71.89	73.56
Liver	12.5k	72.97	78.14	77.58	79.26	79.81	80.23
Aorta	11.8k	48.90	52.44	54.04	59.26	62.56	65.93
Spinal canal	2.4k	76.39	79.17	83.33	90.28	90.28	91.67
Gallbladder	2.4k	19.10	32.58	25.84	42.70	39.33	52.81
Clavicle	1.2k	64.29	89.29	75.00	96.43	96.43	100.00
Ascending aorta	1.6k	23.73	48.28	37.29	56.90	50.85	65.52
Pulmonary artery	1.6k	18.18	28.79	31.82	50.00	53.03	68.18
Breast	1.1k	54.17	73.91	75.00	78.26	75.00	91.30
Pancreas	0.8k	20.51	48.72	38.46	61.54	56.41	74.36
Stomach	0.8k	33.33	54.17	45.83	75.00	79.17	95.83
Average	/	43.85	55.23	54.44	66.29	67.09	76.12

image-caption pairs; and **CT-CLIP** [6], a vision-language foundation model for Chest CT images pre-trained on CT-RATE [6]. Note that none of these baselines support conditional image retrieval, thus we evaluate their performance using retrieval results derived from holistic image and text features across all tasks.

5.2 Metric

Recall@ k evaluates whether the correct items are in the top- k predictions. In image-report retrieval, we consider the paired data as the correct item; In image-image retrieval, we view candidates with similarity over 0.9 as correct items.

NDCG@ k evaluates the predicted ranking by comparing it with the ideal ranking. First, the Discounted Cumulative Gain (DCG) of a ranking is calculated as $DCG = \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}$, where rel_i represents the ground-truth similarity score of the item ranked at position i , and k is the number of items to consider in the ranking. The NDCG is defined as the ratio of the DCG of a predicted ranking to the DCG of the ideal ranking (IDCG) obtained by sorting the items by ground-truth similarity score: $NDCG = DCG/IDCG$.

5.3 Results on Unconditional Retrieval

As demonstrated in Table 1, after fine-tuning, RadIR consistently exceeds the state-of-the-art CLIP models on image-image retrieval task, and on both CXR

and Chest CT datasets. Notably, RadIR can also be applied for image-report retrieval and achieves notable improvement over baselines. These results highlight that RadIR can perform effectively in these traditional retrieval tasks.

5.4 Results on Conditional Image Retrieval

Table 2 shows that RadIR outperforms baselines in 9 out of 10 anatomical regions on CXR images, and achieves the best performance on average; While in Table 3, RadIR consistently outperforms CT-CLIP on metrics. In addition, we observe that RadIR performs better on tail anatomies less frequently mentioned in the report. We hypothesize that this is because baselines trained on image-text pairs exhibit a bias towards more frequent anatomies. While RadIR, supporting conditional retrieval, can effectively adapt its focus to the queried anatomy, demonstrating superior robustness and versatility.

6 Conclusion

In this paper, we propose a novel methodology that leverages dense radiology reports to define image-wise similarity ordering at multiple granularities in a scalable and fully automatic way. We contribute two comprehensive datasets, MIMIC-IR and CTRATE-IR, with comprehensive and fine-grained image similarity ranking annotations for Chest X-ray and CT images. We build RadIR-CXR and RadIR-ChestCT, which demonstrate superior performance in diverse retrieval tasks, and could meet clinical demands flexibly by supporting fine-grained image retrieval conditioned on anatomy.

Acknowledgments. This study was funded by National Key R&D Program of China (No. 2022ZD0161400).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abacha, A.B., Santamaria-Pang, A., Lee, H.H., Merkow, J., Cai, Q., Devarakonda, S.T., Islam, A., Gong, J., Lungren, M.P., Lin, T., et al.: 3d-mir: A benchmark and empirical study on 3d medical image retrieval in radiology. arXiv preprint arXiv:2311.13752 (2023) 2
2. Choe, J., Hwang, H.J., Seo, J.B., Lee, S.M., Yun, J., Kim, M.J., Jeong, J., Lee, Y., Jin, K., Park, R., et al.: Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. *Radiology* **302**(1), 187–197 (2022) 1
3. Delbrouck, J.B., Chambon, P., Chen, Z., Varma, M., Johnston, A., Blankemeier, L., Van Veen, D., Bui, T., Truong, S., Langlotz, C.: Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In: Findings of the Association for Computational Linguistics (2024) 4

4. Denner, S., Zimmerer, D., Bounias, D., Bujotzek, M., Xiao, S., Kausch, L., Schader, P., Penzkofer, T., Jäger, P.F., Maier-Hein, K.: Leveraging foundation models for content-based medical image retrieval in radiology. *arXiv preprint arXiv:2403.06567* (2024) [2](#)
5. Dubey, S.R.: A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(5), 2687–2704 (2021) [1](#)
6. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Doga, I., Durugol, O.F., Dai, W., Xu, M., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834* (2024) [2](#), [4](#), [8](#)
7. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv e-prints* pp. *arXiv-2404* (2024) [1](#)
8. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. pp. 84–92 (2015) [6](#)
9. Hu, B., Vasu, B., Hoogs, A.: X-mir: Explainable medical image retrieval. In: *Winter Conference on Applications of Computer Vision*. pp. 440–450 (2022) [2](#)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2019) [7](#)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* **6**(1), 317 (2019) [2](#), [4](#), [7](#)
12. Kobayashi, K., Gu, L., Hataya, R., Mizuno, T., Miyake, M., Watanabe, H., Takahashi, M., Takamizawa, Y., Yoshida, Y., Nakamura, S., et al.: Sketch-based medical image retrieval. *arXiv preprint arXiv:2303.03633* (2023) [2](#)
13. Lee, H.H., Santamaria-Pang, A., Merkow, J., Oktay, O., Pérez-García, F., Alvarez-Valle, J., Tarapov, I.: Region-based contrastive pretraining for medical image retrieval with anatomic query. *arXiv preprint arXiv:2305.05598* (2023) [2](#)
14. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: *Medical Image Computing and Computer-Assisted Intervention*. pp. 525–536 (2023) [7](#)
15. Müller, H., Michoux, N., Bandon, D., Geissbühler, A.: A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International journal of medical informatics* **73**(1), 1–23 (2004) [1](#)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018) [6](#)
17. Qayyum, A., Anwar, S.M., Awais, M., Majid, M.: Medical image retrieval using deep convolutional neural network. *Neurocomputing* **266**, 8–20 (2017) [1](#)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning*. pp. 8748–8763 (2021) [5](#)
19. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. p. 3876 (2022) [7](#)

20. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463 (2023) [1](#)
21. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), A10a2400640 (2025) [7](#)
22. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Development of a large-scale medical visual question-answering dataset. Communications Medicine **4**(1), 277 (2024) [1](#)
23. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 15004–15019 (2024) [2](#), [4](#)