Advancing Medical Representation Learning Through High-Quality Data

Negin Baghbanzadeh[1,2*], Adibvafa Fallahpour[2,3,5*‡], Yasaman Parhizkar[1,2*], Franklin Ogidi[2], Shuvendu Roy[2,4], Sajad Ashkezari[1,2], Vahid Reza Khazaie[2], Michael Colacci[3], Ali Etemad[4], Arash Afkanpour[2†], and Elham Dolatabadi[1,2†]

[1]York University
[2]Vector Institute
[3]University of Toronto
[4]Queen's University
[5]University Health Network

Abstract. Despite the growing scale of medical Vision-Language datasets, the impact of dataset quality on model performance remains underexplored. We introduce Open-PMC, a high-quality medical dataset from PubMed Central, containing 2.2 million image-text pairs, enriched with image modality annotations, subfigures, and summarized in-text references. Notably, the in-text references provide richer medical context, extending beyond the abstract information typically found in captions. Through extensive experiments, we benchmark Open-PMC against larger datasets across retrieval and zero-shot classification tasks. Our results show that dataset quality-not just size-drives significant performance gains. We complement our benchmark with an in-depth analysis of feature representation. Our findings highlight the crucial role of data curation quality in advancing multimodal medical AI. We release Open-PMC, along with the trained models and our codebase.

Keywords: Multimodal Learning · Representation Learning · Contrastive Learning · Image Decomposition

1 Introduction

In the general domain, Vision-Language (VL) modeling has leveraged massive-scale unlabeled image-text pairs to learn useful representations for a wide range of downstream tasks [19,10]. The medical domain has also seen a surge in efforts to curate extensive medical image-text datasets, often relying on automated crawling of scientific articles [24,13,18,16]. However, the amount of available data in the medical domain remains significantly smaller compared to the general domain. While increasing the volume of data is one way to train high-performance models,

* Equal Contribution
† Equal Advising, arash.afkanpour@vectorinstitute.ai, edolatab@yorku.ca
‡ Work completed during internship at Vector Institute, currently at Cohere

improving data quality remains an under-explored yet promising direction to enhance model performance.

Recent studies [17] highlight that high-quality datasets can mitigate pretraining limitations and enhance model performance, as evidenced in recent work from DeepSeek [15]. Yet, in medical AI, automated extraction of figures and corresponding captions from scientific literature introduces quality challenges. Unlike structured medical reports, such as radiology reports that provide detailed anatomical descriptions, figures in scientific articles are often compound images with captions that are brief or lacking essential clinical context. For example, a dermatology report may provide a detailed description of lesion texture and color variation while a figure caption in a scientific article may simply state 'Example of skin lesion', providing little context for pretraining medical VL models.

In this paper, we investigate how the trade-off between *curation quality and data quantity* in the medical domain impacts learned representations. To address this question, we introduce Open-PMC, a carefully curated image-text medical dataset, and conduct a comprehensive set of experiments. Specifically, we investigate how image decomposition—where compound figures are replaced with subfigures—and contextually enriched captions improve representation learning. Our contributions are summarized as follows:

1. We present Open-PMC, a high-quality dataset extracted from PubMed Central articles (PMC's Open Access Subset), consisting of 2.2 million paired image-text pairs. Each pair includes *subfigures* as images and *captions*, *subcaptions*, and *summarized in-text references* as text. In addition, all pairs are annotated with imaging modalities, including Radiology, Microscopy, and Visible Light Photography (VLP).
2. Through extensive experiments across three imaging modalities, along with ablation studies, we demonstrate that careful dataset curation improves representation learning, leading to better downstream performance.
3. We publicly release Open-PMC along with the trained models and our codebase, providing the research community with valuable resources for advancing AI in medical imaging.

2 Related Work

General-domain multimodal models [19,10,3] owe their success to large-scale datasets, sparking interest in curating medical multimodal datasets, which are typically sourced from PMC articles. ROCO [18] is an open-access dataset which contains approximately 80,000 radiology and 6,000 non-radiology images along with their captions, keywords and other metadata. Lin et al. introduced PMC-OA [14] with 1.6 million image-text pairs. They provided a pipeline for extracting image-text pairs to minimize human involvement. We adapt and expand their pipeline for curating our dataset. More recently, PMC-15M [24] was introduced as a large-scale image-text dataset with 15 million pairs. At the time of writing this paper, however, the dataset has not been released publicly. Biomedica [16]

Table 1: Comparison of medical paired image-text datasets.

Dataset	Size (M)	In-text Ref	Summ Refs	Subfigures	Medical Only	Modality	Open Access
ROCO [18]	0.08	✗	✗	✓	✓	✓	✓
PMC-OA [14]	1.6	✗	✗	✓	✓	✗	✓
PMC-15M [24]	15	✗	✗	✗	✗	✗	✗
Biomedica [16]	24	✓	✗	✗	✗	✓	✓
Open-PMC	2.2	✓	✓	✓	✓	✓	✓

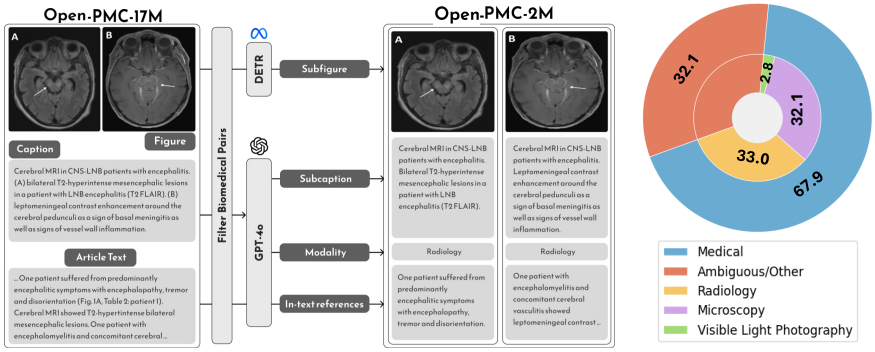

Fig. 1: **Left** Open-PMC-17M comprises 16.7 million image-caption pairs, which undergo rigorous quality curation to produce Open-PMC, including 2.2 million image-text pairs; images are medical subfigures, and texts are captions enriched with both the actual and summarized content of in-text references. **Right** The distribution (%) of each medical image modality within Open-PMC.

is the most recent dataset comprising of 24 million image-text pairs. Using a combination of pretrained image encoders, clustering, and expert annotations they categorized the images with global and local taxonomies, providing modality information for images. Both PMC-15M and Biomedica contain a significant number of non-medical images (plots, charts, etc.). If used for training, this data could hinder model performance in medical applications. In contrast, we have only included high-quality medical images in our dataset. In addition, both PMC-15M and Biomedica contain raw, compound images. Our work builds on these efforts by focusing on high-quality medical images, exploring impact of image decomposition and contextual text augmentation on VL model performance. Table 1 provides a comparative overview of existing image-text datasets, highlighting differences in size and key features.

3 Open-PMC: Curation and Processing

Open-PMC (Fig. 1) is a fine-grained and open-source dataset of 2.2 million high-quality medical image-text pairs. Each example includes: (1) a medical image (subfigure) extracted from an article, (2) its corresponding caption, (3) an

in-text reference from the article body, (4) a *summary* of this reference, and (5) the medical *modality* of the image. To construct Open-PMC, we extended the PMC-OA pipeline [13], integrating additional processing steps, including in-text reference extraction and image modality classification using GPT-4o [8].

3.1 Data Collection and Preprocessing

Our pipeline leveraged *Build-PMC-OA* [13], processing over four million open-access articles from PMC's Open Access Subset (as of June 18, 2024). We extracted figures, captions, and in-text references, employing XML parsing and regular expressions to link figure-caption pairs with their provenance (PMID and PMC-ID).

Quality Control and Filtering We applied a multi-step filtering process, first removing articles with incorrectly formatted XML, missing captions, or syntax errors, yielding 16.7 million image-caption pairs, Open-PMC-17M. We then excluded pairs without predefined medical keywords, reducing the dataset to $880,294$ pairs.

Compound Image Decomposition We used a DEtection TRansformer (DETR)-based model [1] trained on a subset of the MedICaT dataset [23] to decompose compound images into $3,929,247$ single subfigures. Following decomposition, we used a ResNet-101 model trained on the DocFigure dataset [11] to classify images and filter out non-medical samples. This process resulted in a final dataset of 2.2 million high-quality medical images, retaining only figures with high confidence scores in the "Medical" category.

Caption Segmentation and Alignment We used GPT-4o to segment full captions into subcaptions for each subfigure. To align captions with subfigures, we leveraged object localization (YOLOv3) and recognition (ResNet-152) from Exsclaim [22,20,6] to detect labels and match them to subcaptions. For subfigures without identifiable labels, we assigned the entire original caption to preserve textual context.

3.2 Textual Augmentation and Contextualization

In-text Reference Extraction and Summarization We extracted paragraphs referencing a figure in the article using XML cross-reference tags and segment them into individual sentences using regular expressions. However, in-text references often span multiple paragraphs, exceeding the model's processing limits, and relevant details may appear outside the sentences that directly reference a target subfigure. Moreover, for compound figures, determining which references correspond to specific subfigures is non-trivial. To address these challenges, we used GPT-4o-mini to generate focused summaries, distilling the most relevant contextual information while preserving critical details.

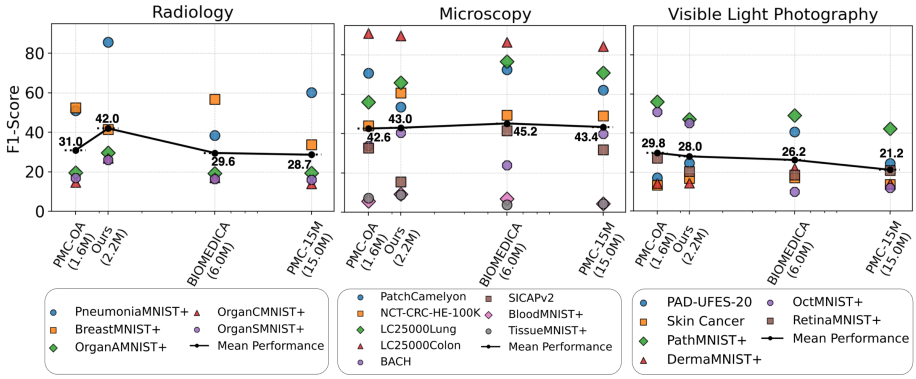

Fig. 2: Zero-shot classification F1-scores across different VL models trained on datasets of varying sizes, evaluated on downstream tasks split by image modality. Each marker represents performance on an individual task, while the solid line indicates the mean performance across all tasks. *Ours* indicates Open-PMC.

Image Modality Assignment We used GPT-4o-mini to extract image modality information from captions, categorizing images as diagnostic or non-diagnostic. Diagnostic images were further classified into Radiology (e.g., X-ray, Ultrasound, CT, or MRI), VLP (e.g., Dermatology, Endoscopy), and Microscopy. To validate accuracy, three reviewers independently assessed 1,000 randomly selected images (one per modality), achieving an overall 82.5% agreement with the automated classification.

4 Experiments

We perform an extensive evaluation of Open-PMC for medical representation learning by training encoders, referred to as VL models, via contrastive learning, which has gained widespread popularity and ubiquity [19,25,21]. We compare the performance of these models across a wide range of downstream tasks against models trained on similar datasets, some larger than Open-PMC (specifically, PMC-15M [24], and Biomedica [16]). Since our primary focus in this paper is on the dataset rather than modeling, we refer to datasets rather than model names across all experiments. We conduct our experiments using the **mmlearn** multimodal learning framework. The code and experimental setup are available at https://github.com/vectorInstitute/pmc-data-extraction/.

4.1 Setup

Pretraining All encoders are trained using a vanilla contrastive loss to align vision and text representations. Our initial encoders comprise PubMedBERT [5]

https://github.com/VectorInstitute/mmlearn

as the text encoder and a ViT-B/16 transformer [2], pretrained on ImageNet, as the vision encoder. For a fair comparison, we train the same encoder architecture on four datasets: PMC-OA [13], ROCO [18], OPEN-PMC, and OPEN-PMC without in-text (w/o in-text) references. The OPEN-PMC encoders are trained for 64 epochs, while training durations for other datasets are adjusted to ensure all models train on the same total number of examples. For each encoder, the best-performing checkpoint is selected based on validation retrieval performance. Models trained on PMC-15M (BiomedCLIP) and BIOMEDICA (BMCA-CLIP$_{\text{CF}}$) were downloaded directly from the corresponding HuggingFace pages.

Table 2: Retrieval performance (Recall@200) of VL models. The last two columns, Average Recall (AR) and Mean Reciprocal Rank (MRR) aggregate the results across all tasks. Highest performance values are in bold, second-best are underlined.

	Text-to-Image			**Image-to-Text**			**Summary**	
Model	**MIMIC-CXR**	**Quilt**	**DeepEyeNet**	**MIMIC-CXR**	**Quilt**	**DeepEyeNet**	**AR**	**MRR**
ROCO	0.080	0.024	0.079	0.084	0.023	0.108	0.066	0.208
PMC-OA	0.139	0.142	<u>0.152</u>	0.152	0.149	<u>0.157</u>	0.149	0.361
PMC-15M	<u>0.162</u>	<u>0.186</u>	0.147	<u>0.185</u>	<u>0.166</u>	**0.162**	<u>0.168</u>	<u>0.556</u>
BIOMEDICA	0.094	**0.195**	0.145	0.076	**0.169**	0.155	0.139	0.506
OPEN-PMC	**0.170**	0.166	**0.183**	**0.189**	0.162	0.147	**0.170**	**0.653**

Downstream Tasks We evaluate encoders on retrieval and zero-shot classification tasks. Retrieval includes both image-to-text (I2T) and text-to-image (T2I) retrieval on three benchmark datasets: Quilt [9] (microscopy), MIMIC-CXR [12] (radiology), and DeepEyeNet [7] (VLP). For classification, we conduct zero-shot evaluations across radiology (5 tasks), microscopy (8 tasks), and VLP (6 tasks).

4.2 Findings

Data Quality over Quantity Despite OPEN-PMC being much smaller (3 to 7 fold) than BIOMEDICA and PMC-15M, the model trained on OPEN-PMC achieves performance that is not only comparable but also superior in certain tasks. Fig. 2 presents zero-shot classification results, excluding models trained on ROCO due to their consistently low performance. The models trained on OPEN-PMC achieve the highest mean performance in radiology, as measured by F1-score, and the least performance variability in VLP. For microscopy classification, it underperforms by only 4.87% relative to the best model, with a confidence range of 2.87% to 6.87%, incorporating an estimated standard deviation of 2%.

Retrieval results are shown in Table 2. OPEN-PMC achieves the best Average Recall (AR) and Mean Reciprocal Rank (MRR) among the datasets while being

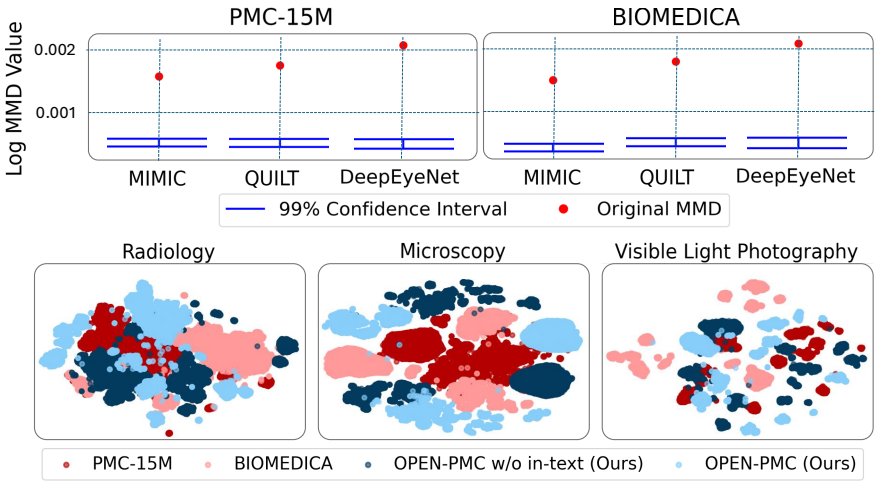

Fig. 3: Comparison of representation spaces of different VL models. **(Top)** MMD values between representations learned from OPEN-PMC versus PMC-15M and BIOMEDICA. Red dots indicate observed MMD values, and blue bars are 99% bootstrap confidence interval of the permutation test. **(Bottom)** t-SNE visualizations of VL models embeddings, illustrating the structure and separation of the learned representation spaces.

smaller than PMC-15M and BIOMEDICA, underscoring the crucial role of data quality in the medical domain. Similar to the classification task, OPEN-PMC consistently delivers the highest recall performance in radiology and the best score for T2I retrieval on DeepEyeNet (VLP).

Distinct Representations vs. Prior Medical Datasets Leveraging VL models trained on OPEN-PMC to learn representations for all three medical imaging modalities as shown in Fig. 3 (bottom three plots), reveals that OPEN-PMC produces a distinct latent structure compared to PMC-15M and BIOMEDICA in the 2D t-SNE space.

To quantify these differences, we employ Maximum Mean Discrepancy (MMD) [4]. Let D denote a dataset (e.g., MIMIC-CXR images), and let $\phi(D)$ and $\psi(D)$ denote the feature embeddings obtained by applying encoders ϕ and ψ on D. For instance, ϕ corresponds to an encoder trained on OPEN-PMC, while ψ represents one trained on BIOMEDICA. To test whether the underlying distributions of $\phi(D)$ and $\psi(D)$ are different, we conduct a permutation test on the MMD values computed between their respective representation sets. We repeat this experiment with $D \in \{\text{MIMIC-CXR}, \text{Quilt}, \text{DeepEyeNet}\}$ for encoders trained on OPEN-PMC, PMC-15M, and BIOMEDICA.

Our MMD analysis reinforces the t-SNE visualization, revealing a substantial difference in the learned representations. As shown in Fig. 3 (top), the MMD

Table 3: Zero-shot classification F1-score comparison between VL models trained on compound images and subfigures for radiology. Performance differences are shown in parentheses.

Model	PneumoniaMNIST+	BreastMNIST+	OrganAMNIST+	OrganCMNIST+	OrganSMNIST+
Compound	63.55	48.07	20.83	18.63	18.60
Subfigures	73.58 (10.03 ↑)	51.47 (3.40 ↑)	26.68 (5.85 ↑)	19.96 (1.33 ↑)	20.89 (2.29 ↑)

Table 4: Retrieval performance (Recall@200) comparison between VL models trained on Open-PMC with and without in-text reference summaries. Performance differences are shown in parentheses, with green indicating higher retrieval performance for Open-PMC.

	Text-to-Image Retrieval			Image-to-Text Retrieval			Summary
Model	MIMIC-CXR	Quilt	DeepEyeNet	MIMIC-CXR	Quilt	DeepEyeNet	AR
Open-PMC w/o in-text	0.165	0.166	0.157	0.183	0.162	0.132	0.160
Open-PMC	**0.170** (0.005 ↑)	0.147 (0.019 ↓)	**0.183** (0.026 ↑)	**0.189** (0.006 ↑)	0.139 (0.023 ↓)	0.147 (0.015 ↑)	**0.162** (0.002 ↑)

values (red dots) between the two representation sets are significantly larger than the 99% bootstrap confidence interval of the permutation test. This trend remains consistent across MIMIC-CXR, Quilt, and DeepEyeNet. These results indicate the impact of Open-PMC in shaping different representations from prior medical datasets.

Ablation Study We hypothesize that Open-PMC's superior performance in radiology tasks stems from the quality of its figures—particularly the use of subfigures instead of compound images—and the inclusion of summarized in-text references added to the caption, both of which are paired with subfigures for contrastive learning. To test this, we conduct two ablation experiments. Since the decomposition pipeline was pre-trained for radiology, our first experiment focuses on radiology classification. We created an alternative version of our dataset where images remained in their original compound form. This reduced the dataset size to 792,000 pairs. To ensure a fair comparison, we randomly sampled an equal number of subfigures from Open-PMC to match the size of the pairs. As shown in Table 3, zero-shot classification performance drops when transitioning from subfigures to compound figures. This is one of the factors that distinguishes Open-PMC from PMC-15M and Biomedica.

The second experiment examines the impact of in-text summaries on retrieval performance. We compared models trained on Open-PMC with one trained on Open-PMC w/o in-text summaries on retrieval tasks (Table 4). As expected, Open-PMC outperforms Open-PMC w/o in-text overall, confirming that incorporating in-text summaries enhances contextual knowledge, strengthening the connection between images and their textual descriptions.

5 Conclusion

Our study highlights the critical role of high quality dataset curation in medical VL learning. By introducing Open-PMC, we demonstrate that image decomposition and incorporating in-text summaries enhance representation learning beyond dataset scale alone. However, a key limitation is that our image decomposition pipeline was primarily optimized for radiology images, limiting its effectiveness for microscopy and other medical modalities. Moreover, further data quality checks and refinement through uncertainty estimation, outlier detection, and human-in-the-loop validation could enhance dataset reliability. Future work will focus on expanding image decomposition techniques for diverse medical imaging modalities and incorporating more robust data quality assurance methods.

Acknowledgments. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (https://vectorinstitute.ai/partnerships/current-partners/). This research was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Canadian Institutes of Health Research (CIHR) Special Call through the Centre for Research on Pandemic Preparedness and Health Emergencies.

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

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
3. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
4. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
5. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing (2020)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), `https://arxiv.org/abs/1512.03385`
7. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al.: Deepopht: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)

8. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
9. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in Neural Information Processing Systems **36** (2024)
10. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
11. Jobin, K., Mondal, A., Jawahar, C.: Docfigure: A dataset for scientific document figure classification. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 1, pp. 74–79. IEEE (2019)
12. 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)
13. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
14. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine p. 102611 (2023)
15. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
16. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Katzer, A.W., Chiu, C., et al.: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. arXiv preprint arXiv:2501.07171 (2025)
17. Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., Schmidt, L.: Quality not quantity: On the interaction between dataset design and robustness of clip. Advances in Neural Information Processing Systems **35**, 21455–21469 (2022)
18. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. pp. 180–189. Springer (2018)
19. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement (2018), `https://arxiv.org/abs/1804.02767`
21. Roy, S., Parhizkar, Y., Ogidi, F., Khazaie, V.R., Colacci, M., Etemad, A., Dolatabadi, E., Afkanpour, A.: Benchmarking vision-language contrastive methods for medical representation learning. arXiv preprint arXiv:2406.07450 (2024)

22. Schwenker, E., Jiang, W., Spreadbury, T., Ferrier, N., Cossairt, O., Chan, M.K.: Exsclaim! - an automated pipeline for the construction of labeled materials imaging datasets from literature. arXiv e-prints pp. arXiv–2103 (2021)
23. Subramanian, S., Wang, L.L., Mehta, S., Bogin, B., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., Hajishirzi, H.: Medicat: A dataset of medical images, captions, and textual references. arXiv preprint arXiv:2010.06000 (2020)
24. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
25. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Li, X., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)