# Open-PMC-18M: A High-Fidelity Large Scale Medical Dataset for Multimodal Representation Learning

Negin Baghbanzadeh[1,2*], Mohammed Saidul Islam[1*], Sajad Ashkezari[1,3*], Elham Dolatabadi[1,2†], and Arash Afkanpour[1†]

[1]Vector Institute
[2]York University
[3]University of Waterloo

**Abstract.** In biomedical vision-language modeling, datasets are typically mined from scientific literature where compound figures are paired with short and often partially informative captions. Prior work on subfigure extraction has been limited in both dataset size and generalizability. In addition, no existing effort has incorporated rich medical context in image-text pairs. We revisit data curation as a foundational component of effective biomedical representation learning. We address this by introducing a scalable curation pipeline that integrates transformer-based subfigure detection, subcaption extraction, and contextual text enrichment to improve alignment quality. We introduce **Open-PMC-18M**, a large-scale high-fidelity biomedical dataset comprising 18 million image-text pairs, spanning radiology, microscopy, and visible light photography. We train vision-language models on our dataset and perform extensive evaluation on 6 retrieval and 19 zero-shot classification tasks across three major modalities. The models trained on our dataset establish new state-of-the-art results in medical representation learning. We publicly release the dataset and model, along with the code.

**Keywords:** biomedical · vision-language models · subfigure extraction · representation learning · data quality.

## 1 Introduction

Quality of representations in vision-language models (VLMs) [22,10,5] is highly sensitive to the fidelity of image-text alignment under contrastive pretraining [18,22]. In natural web data, such as those used for CLIP [22] and subsequent models [17], the paired text (captions, metadata, or surrounding paragraphs) is

* Equal Contribution
† Equal Advising, edolatab@yorku.ca, arash.afkanpour@vectorinstitute.ai
https://huggingface.co/collections/vector-institute/open-pmc-18m
https://github.com/vectorInstitute/pmc-data-extraction

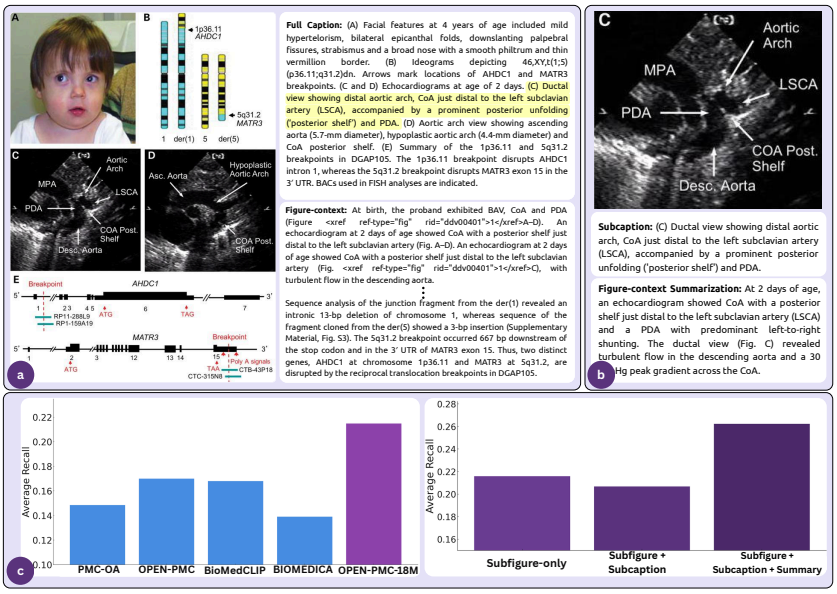


**Fig. 1.** *(a)* Compound figure and the corresponding full caption (subcaption for subfigure 'C' is highlighted) and in-text reference related to the figure from BIOMEDICA dataset; *(b)* Extracted subfigure image from OPEN-PMC-18M, and the corresponding extracted subcaption and summary of in-text reference; *(c)* Average retrieval performance: our model vs. other models (**left**), and retrieval results on MIMIC-CXR of model versions trained in different settings: *subfigure only* vs. *subfigure + subcaption* vs. *subfigure + subcaption + summary* (**right**). The compound image in *(a)* is originally from [20].

typically descriptive and diverse, providing rich linguistic grounding for visual concepts [24,4]. In the biomedical domain, however, the paired supervision is constructed differently.

Most medical VLMs are trained on datasets mined from scientific literature, such as PubMed Central (PMC)(e.g, PMC-15M [25] and BIOMEDICA [16]), in which figures and captions are extracted to form image-text pairs (see Figure 1(a)). This introduces two fundamental challenges. First, biomedical figures are often compound or multi-panel, combining heterogeneous content, including radiology scans, microscopy images, plots, and annotations. Second, the corresponding captions are frequently symbolic, abbreviated, or context-dependent, relying heavily on in-text references and not self-contained descriptions. Pairing compound figures with partially informative captions yields coarse image-text alignments that degrade representation quality. Such mismatches encourage models to rely on superficial textual cues that weaken cross-modal correspondence and limit transfer to downstream tasks.

https://pmc.ncbi.nlm.nih.gov/

To our knowledge, no prior work has jointly addressed image and text enrichment in the biomedical domain. A few prior works tackled subfigure extraction but at limited scale and accuracy [19,13,1]; contextual enrichment has been largely unaddressed. This raises an important gap in the field: *How does subfigure extraction and contextual text summarization affect the quality of learned representations in the medical domain, particularly given the known sensitivity of contrastive objectives to dataset alignment and scale during pretraining?*

We introduce Open-PMC-18M, the first large-scale and carefully curated dataset comprising 18 million subfigure-text pairs, where each text entry consists of an associated *subcaption*, extracted from the original caption, and a summary of inline context that explicitly references the subfigure (Figure 1 (b)). Starting from the Biomedica corpus [16], which is extracted from PubMed Central, we apply metadata-level filtering using label annotations and a ResNet-based classifier to remove non-medical or schematic content. For subfigure extraction, we train a high-performance detector on a dataset of 500,000 programmatically generated compound figures. For subcaption extraction we utilized Qwen2.5-VL-32B-Instruct VLM [2] and for text summarization we used Qwen2.5-14B-Instruct [21].

Our main contributions are as follows:

- We curate and release Open-PMC-18M, a large-scale high-fidelity biomedical image-text dataset with 18 million image-text pairs, each consists of a subfigure paired with the corresponding subcaption and image-context summary.
- We propose a scalable and accurate subfigure extraction pipeline based on transformer-based object detection, trained on 500,000 compound figures, achieving state-of-the-art performance on ImageCLEF 2016 [12,7] and synthetic evaluation sets.
- We perform a comprehensive evaluation of VLMs trained on Open-PMC-18M, demonstrating improved performance in retrieval, classification, and robustness across multiple medical benchmarks in radiology, microscopy, and visible light photography (VLP).

## 2 Open-PMC-18M Curation

Our dataset curation pipeline (Figure 2(a)) contain three key stages: *(i)* Initial Figure Collection and Filtering, *(ii)* Vision-Based Subfigure Extraction, and *(iii)* Textual Enrichment via subcaption extraction and summary generation. Below we describe each stage in detail.

### 2.1 Initial Figure Collection and Filtering

To curate our dataset, we start with the Biomedica dataset [16]. Biomedica contains approximately 24 million image-caption pairs along with in-text figure references, often spanning multiple paragraphs which are extracted from

**Table 1.** Comparison of subfigure detection performance on two datasets.

| Model | Holdout Set. | | ImageCLEF 2016 | |
|---|---|---|---|---|
| | mAP (%) | F1 (%) | mAP (%) | F1 (%) |
| MedICaT-trained DETR | 33.22 | 73.18 | 28.20 | 64.85 |
| Qwen2.5-VL-32B-Instruct | 40.31 | 79.01 | 30.12 | 68.12 |
| Ours (DAB-DETR) | **98.58** | **99.96** | **36.88** | **73.55** |

PubMed Central Open Access. To improve dataset quality, we apply a filtering step and retain only those pairs primarily categorized as clinical imaging, microscopy, immunoassays, or chemical structure. This yields a dataset of 6 million pairs, which we refer to as PMC-6M in this paper.

### 2.2 Vision-Based Subfigure Extraction

To enable accurate large-scale subfigure extraction from biomedical compound figures, we train a transformer-based object detector using the DAB-DETR architecture [14], which improves localization and convergence via dynamic anchor queries. Unlike prior work [13], which was trained on only 2,069 manually annotated MedICaT figures [23], we construct and train on a large-scale dataset of 500,000 compound figures, the first of its kind in the biomedical domain.

To train a subfigure extraction model, we generate a synthetic dataset by reversing the subfigure extraction process: we programmatically construct compound images by composing multiple single-panel biomedical images using predefined layouts as illustrated in Figure 2(b). Layout diversity is achieved through randomized grid structures, margins, labeling schemes, and aspect ratios. Panels are sampled from radiology, microscopy, and VLP datasets sourced from well-known benchmark datasets, including ROCO, SICAP, HAM10000, MedMNIST variants, PAD-UFES-20, and PlotQA, with balanced modality representation and mixed-modality compositions to promote generalization as shown in Figure 2(d).

The object detection model is then trained on this dataset for 40 epochs (batch size 64, learning rate $10^{-5}$) and evaluated on both a holdout set with 20,000 images and the ImageCLEF 2016 compound figure benchmark [12,7]. Compared to a DETR model trained on MedICaT and a zero-shot adapted Qwen2.5-VL-32B-Instruct model [2], our DAB-DETR achieves superior performance as reported in Table 1. Figure 2(c) demonstrates examples of subfigure extraction using our model across heterogeneous layouts and imaging modalities from the ImageCLEF dataset.

### 2.3 Text Enrichment

To enhance semantic alignment for each subfigure, we augment our dataset with (1) automatically extracted subcaptions and (2) summarized in-text references.

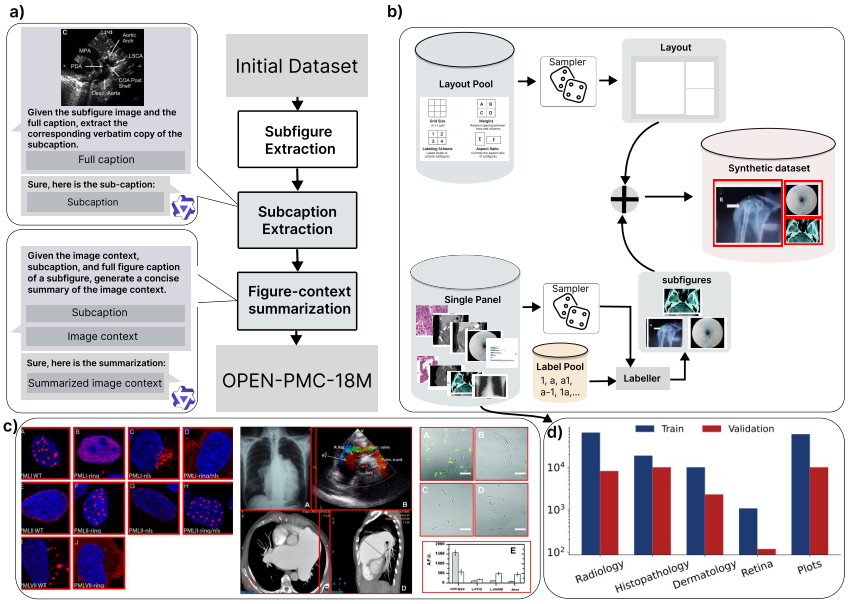


**Fig. 2.** *(a)* Overview of the Open-PMC-18M construction pipeline. *(b)* The Subfigure Extraction pipeline for creating compound figures that are used to train the DAB-DETR model. *(c)* Example subfigures extracted using the DAB-DETR model from the ImageCLEF dataset. *(d)* Distribution of the modality of single panel images used to construct the compound figure corpus.

The majority of captions in the dataset consist of multiple subcaptions, generally distinguished by delimiters such as (a), (b), or (A–F). As such, we use Qwen2.5-VL-32B-Instruct [2] to extract subcaptions from the full caption. To ensure determinism and prevent hallucinations or inaccuracies, the model was explicitly instructed to extract a verbatim copy of the subcaption directly from the full caption, without adding any additional text or explanation. Manual evaluation of 1,000 samples shows 94% extraction accuracy, with errors largely due to missing subcaptions; 13.28% of cases contain no distinct subcaption and thus default to the full caption. Inference was conducted on 32×A100 GPUs (4 nodes), requiring roughly 13,474 GPU-hours (17.5 wall-clock days).

To further enrich contextual information, we leverage inline text references mentioning each figure. These references frequently extend across multiple paragraphs. In many cases, important contextual details appear outside the sentences that directly correspond to a subfigure, making it difficult to determine which pieces are related to a particular subfigure. To mitigate these issues, we employ the Qwen2.5-14B-Instruct [21] model to generate concise summaries that distill the most relevant information and essential details for each subfigure.

Decomposing the compound images of PMC-6M with our DAB-DETR model produces 32M subfigure-caption pairs. Filtering by modality metadata (Clinical

Image and Microscopy only) reduces this to 26M pairs, and a final medical-relevance screening using a ResNet-101 classifier [13] yields 18M high-quality image-text pairs. Open-PMC-18M comprises 73% microscopy, 18% radiology, 8% VLP, and 1% other clinical images. Full captions are variable (avg. 165.8 tokens; 19.5% >256 tokens), whereas subcaptions are short and consistent (avg. 30 tokens; 83% <50 tokens). Context summaries are moderately longer (avg. 54 tokens) and capped at 256 tokens. Most compound figures contain 1-3 subfigures.

## 3 Experiments

### 3.1 Setup

We pretrain vision and text encoders using a standard image-text contrastive objective [18]. The vision backbone is ViT-B/16 [3] (ImageNet pretrained) and the text encoder is PubMedBERT [6]. Models are trained for 60 epochs with a batch size of 2,048 on image-text pairs consisting of subfigures, extracted subcaptions, and summarized context, using 8×A100 GPUs. Checkpoints are selected based on cross-modal retrieval performance on a 50k holdout validation set.

To assess the impact of dataset scale and curation quality, we train all models under a unified architecture and protocol for controlled comparison. Baselines include models trained on PMC-6M (compound-level pairs), a reproduced PMC-OA setup [13], and publicly available checkpoints from PMC-15M [25] and Biomedica [16] (BMC-CLIP$_{\text{CF}}$). All external models are evaluated under the same downstream retrieval protocols to ensure fair comparison.

We evaluate encoder performance on cross-modal retrieval and classification. Retrieval (image-to-text and text-to-image) is assessed on Quilt [9] (microscopy), MIMIC-CXR [11] (radiology), and DeepEyeNet [8] (VLP). Robustness in retrieval is measured under standard visual perturbations (brightness, shift, rotation, flip, zoom) following [15], with statistical significance assessed via the Wilcoxon signed-rank test ($p < 0.01$). We also conduct zero-shot classification across 19 tasks spanning radiology, microscopy, and VLP, encoding images and text using the pretrained encoders.

### 3.2 Downstream Tasks

**Cross-Modal Retrieval and Robustness.** Tables 2 and Figure 1(c)(left) report cross-modal retrieval results (Recall@200) across Quilt, MIMIC-CXR, and DeepEyeNet. Models trained on Open-PMC-18M and PMC-6M consistently outperform PMC-15M and Biomedica in both retrieval directions, with Open-PMC-18M achieving a new state-of-the-art Average Recall (AR) of 21.64, representing a 27% relative improvement over the strongest prior model. Robustness, measured as the ratio of performance under standard visual perturbations to original performance as shown in Figure 3, further shows that Open-PMC-18M-trained models maintain superior stability, with statistically significant gains ($p < 0.01$) on Quilt and DeepEyeNet, highlighting improved generalization under heterogeneous imaging conditions.

**Table 2.** Retrieval performance (Recall@200) of all models trained on paired image-caption pairs in the medical domain. The last column, Average Recall (AR), aggregates results across all tasks. Highest values are in bold, second-best are underlined. PMC-6M is trained on compound figures from a filtered subset of BIOMEDICA without subfigure decomposition. The BIOMEDICA checkpoint is obtained from HuggingFace.

| | Image-to-Text | | | Text-to-Image | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **MIMIC-CXR** | **Quilt** | **DeepEye Net** | **MIMIC-CXR** | **Quilt** | **DeepEye Net** | **AR** |
| PMC-OA | 0.139 | 0.142 | 0.152 | 0.152 | 0.149 | 0.157 | 0.148 |
| OPEN-PMC | 0.170 | 0.166 | <u>0.183</u> | 0.189 | 0.162 | 0.147 | 0.170 |
| BioMedCLIP | 0.185 | 0.165 | 0.162 | 0.162 | 0.185 | 0.146 | 0.167 |
| BIOMEDICA | 0.076 | 0.169 | 0.155 | 0.093 | 0.195 | 0.145 | 0.139 |
| PMC-6M | **0.250** | <u>0.203</u> | 0.172 | **0.257** | <u>0.220</u> | <u>0.170</u> | <u>0.212</u> |
| OPEN-PMC-18M | <u>0.215</u> | **0.226** | **0.192** | <u>0.210</u> | **0.256** | **0.197** | **0.216** |

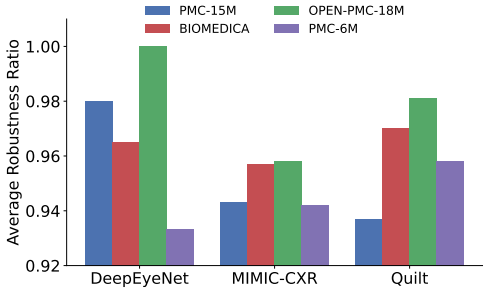


**Fig. 3. Robustness across retrieval benchmarks.** Performance is measured as the ratio of performance under visual perturbations to original performance (higher is better).

**Zero-shot Classification.** Zero-shot classification results (Table 3) show that models trained on OPEN-PMC-18M achieve the highest overall average across modalities, with particularly strong gains in microscopy and radiology. Across 19 tasks spanning radiology, microscopy, and VLP, OPEN-PMC-18M ranks first in 13 and second in 1, outperforming existing biomedical VLMs.

## 4 Effects of Fine-Grained & Context-Enriched Language Supervision on Medical Representation Learning

We perform an ablation to examine the role of increasing textual fidelity. The results of this analysis are summarized in Figure 1(c)(right). To make the experiment computationally feasible, we restrict training to the radiology subset of OPEN-PMC-18M, which contains 3.2 million image-text pairs. All model vari-

**Table 3. Zero-shot classification average F1-scores (%).** Results across radiology, vision-language pathology (VLP), and microscopy.

| Model | Radiology | VLP | Microscopy | Average |
|---|---|---|---|---|
| PMC-OA | 30.95 | 29.79 | 42.62 | 34.45 |
| Open-PMC | **36.19** | 28.30 | 39.12 | 34.54 |
| BioMedCLIP | 28.62 | 21.17 | 43.35 | 31.04 |
| BIOMEDICA | 29.56 | 26.21 | 45.16 | 33.64 |
| PMC-6M | 30.28 | <u>31.50</u> | <u>46.35</u> | <u>36.04</u> |
| Open-PMC-18M | <u>34.55</u> | **32.10** | **55.35** | **39.77** |

ants use the same encoder architecture, optimization settings, and contrastive pretraining objective.

In the *Subfigure-only* variant, each subfigure is paired with its full caption. In *Subfigure + Subcaption*, the full caption is replaced by the extracted subcaption corresponding to that subfigure, providing finer-grained text alignment. Finally, in *Subfigure + Subcaption + Summary*, each subcaption is further augmented with a summary of its in-text context, capturing additional diagnostic or experimental details.

Results in Figure 1 (c)(right) on MIMIC-CXR show that replacing full captions with subcaptions alone does not yield a performance gain. In fact, the reduced textual richness leads to a slight drop in AR, highlighting that subcaptions, while more locally precise, may be too sparse to support robust alignment. However, when contextual summaries are added, performance increases substantially, indicating that the combination of panel-specific subcaptions and enriched contextual information provides the strongest supervision signal.

## 5 Conclusion, Responsible Use, and Future Implications

In this work, we addressed a fundamental limitation in current biomedical vision-language pretraining: the reliance on misaligned compound figures and weak medical context in image-text pairs. We introduced Open-PMC-18M, a large-scale, high-fidelity dataset constructed through accurate subfigure extraction and contextual text enrichment. Our systematic evaluations demonstrate that models trained on Open-PMC-18M achieve consistent gains across radiology, microscopy, and VLP, outperforming existing PMC-derived datasets in retrieval, classification, and robustness. Our findings suggest a broader principle that, in biomedical VLMs, data quality and dataset scale should be viewed as complementary dimensions.

While our models are not intended for clinical deployment, they may be fine-tuned for downstream applications. Any such use, however, requires rigorous external validation, careful assessment of failure modes, and adherence to clinical safety standards. Without these safeguards, real-world deployment could introduce significant risks. Finally, because Open-PMC-18M is derived from

open-access scientific literature, it may reflect structural biases related to publication practices, institutional representation, imaging protocols, and population demographics, which could limit generalizability. Future work should focus on targeted evaluation, dataset diversification, and bias-aware development.

**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. Baghbanzadeh, N., Fallahpour, A., Parhizkar, Y., Ogidi, F., Roy, S., Ashkezari, S., Khazaie, V.R., Colacci, M., Etemad, A., Afkanpour, A., et al.: Advancing medical representation learning through high-quality data. arXiv preprint arXiv:2503.14377 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. 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)
4. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)
5. 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)
6. 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)
7. García Seco de Herrera, A., Schaer, R., Bromuri, S., Müller, H.: Overview of the ImageCLEF 2016 medical task. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (September 2016)
8. 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)

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. 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)
12. Kalpathy-Cramer, J., García Seco de Herrera, A., Demner-Fushman, D., Antani, S., Bedrick, S., Müller, H.: Evaluating performance of biomedical image retrieval systems– an overview of the medical image retrieval task at ImageCLEF 2004–2014. Computerized Medical Imaging and Graphics (2014)
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. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: International Conference on Learning Representations (2022)
15. Liu, X., Li, W., Yuan, Y.: Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–66. Springer (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. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European conference on computer vision. pp. 529–544. Springer (2022)
18. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): a multimodal image dataset. In: CVII-STENT/LABELS@MICCAI. pp. 180–189. Springer (2018)
20. Quintero-Rivera, F., Xi, Q.J., Keppler-Noreuil, K.M., Lee, J.H., Higgins, A.W., Anchan, R.M., Roberts, A.E., Seong, I.S., Fan, X., Lage, K., et al.: Matr3 disruption in human and mouse associated with bicuspid aortic valve, aortic coarctation and patent ductus arteriosus. Human molecular genetics **24**(8), 2375–2389 (2015)
21. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), https://arxiv.org/abs/2412.15115
22. 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)

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. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data (2024), https://arxiv.org/abs/2309.16671
25. 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)