PedCLIP: A Vision-Language model for Pediatric X-rays with Mixture of Body part Experts

Ta Duc Huy[1(✉)], Abin Shoby[1], Sen Tran, Yutong Xie[5], Qi Chen[1], Phi Le Nguyen[4], Akshay Gole[1], Lingqiao Liu[1], Antonios Perperidis[1], Mark Friswell[3], Rebecca Linke[3], Andrea Glynn[3], Minh-Son To[2], Anton van den Hengel[1], Johan Verjans[1], Zhibin Liao[1], and Minh Hieu Phan[1]

[1] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, Australia tdh512194@gmail.com
[2] Flinders University, Adelaide, Australia
[3] Woman's and Children's Hospital, Adelaide, Australia
[4] Hanoi University of Science and Technology, Hanoi, Vietnam
[5] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Abstract. Vision-language models have demonstrated remarkable success in general medical image analysis, yet their application in pediatric imaging remains significantly underexplored. These models show limited performance on pediatric datasets, primarily due to domain gaps stemming from anatomical differences, lower radiation doses, and pediatric-specific diseases. To this end, we present the first pediatric vision-language pre-training framework, dubbed PedCLIP, trained on a comprehensive pediatric imaging dataset comprising 404,670 X-rays of pediatric patients across diverse anatomical regions. To address anatomical diversity, we introduce a Mixture of Body part Experts design, with each expert specializing in learning features from distinct anatomical regions. Experimental evaluation across eleven downstream tasks demonstrates that our model significantly outperforms current state-of-the-art vision-language models, achieving superior diagnostic accuracy in challenging pediatric conditions, including rare diseases such as pediatric inflammatory arthritis. Code is available: https://github.com/tadeephuy/PedCLIP

Keywords: pediatric · mixture-of-experts · vision-language-model.

1 Introduction

Pediatric imaging serves as a critical tool for early diagnosis and substantially influences long-term health outcomes in children. Beside specialized medical models [22,9,21,20,7], recent advances in medical foundational vision-language models (VLMs) [26,23,2,27,8,19,25,13] have attracted significant attention due to their robust performance across various downstream tasks. However, their application to pediatric data remains limited.

These pretrained models have demonstrated suboptimal performance on pediatric datasets, indicating a clear domain gap (Fig. 1). In this study, we explore

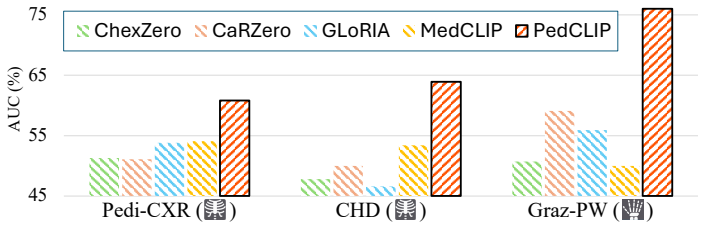

Fig. 1. Zero-shot classification performance of **adult Chest** X-ray pretrained models (ChexZero [23], CaRZero [13], GLoRIA [6], MedCLIP [26]) versus our **pediatric multi-body-parts** pretrained model. *Adult* Chest vision language models demonstrate subpar performance on *Pediatric* Chest (Pedi-CXR [18], CHD-CXR [28]) and other body part (Wrist) [16] X-rays datasets.

the underlying factors contributing to this gap. First, pediatric X-rays show anatomical differences (e.g., smaller lung fields, developing rib cage) compared to adult . Second, pediatric imaging protocols employ lower radiation doses [24], affecting image property. Third, most existing medical foundation VLMs are predominantly trained on chest X-ray image-text pairs dataset from MIMIC [11]. This results in a bias toward the chest region due to the lack of data samples collected from other body parts. Consequently, these models underperform when applied to other anatomical regions (i.e., hip, knee) (see Fig. 1).

To bridge this gap, we propose the *first-of-its-kind* foundational pediatric VLMs. Our model is pretrained on 400K pediatric X-rays collected from 200K pediatric cases. This dataset spans multiple anatomical regions to support the model's generalizability across body parts. First, such a dataset exhibits a heterogeneity property due to the imaging differences for different body parts. We empirically show that training a standard VLM on multiple body parts results in subpar performance, which we attribute to *feature interference* when learning heterogeneous anatomical features. To address these issues, we propose a Mixture of Body Experts (MoBE) architecture, where each expert specializes in a body part, allowing the model to seamlessly handle the anatomical diversity. The experimental results demonstrate that the MoBE design alleviates *feature interference* when the VLM is trained on multiple body part data.

Our contributions are summarized as follows: (i) We introduce the *first* foundational pediatric VLM, dubbed PedCLIP, pretrained on a diverse pediatric dataset encompassing multiple anatomical regions; (ii) We employ Mixture of Body Experts, wherein each expert specializes in learning features from a distinct body part, effectively addressing the heterogeneity of the dataset. (iii) We demonstrate state-of-the-art performance across five tasks on five pediatric datasets, outperforming existing VLMs in both zero-shot and supervised settings. We will open source our foundational PedCLIP model to facilitate open pediatric imaging analysis research.

2 Method

In this section, we first introduce the dataset curation process and then describe the foundational model with Mixture of Body part Experts (MoBE).

2.1 Pediatric X-rays (PeXR) dataset curation

To pre-train our foundational pediatric VLM, we present PeXR, a large-scale dataset comprising 404,670 pediatric X-rays of 23 anatomical regions collected from 262,577 imaging studies from 94,543 pediatric patients. Major anatomical regions include chest, wrist and forearms take up 46%. Each case includes X-rays filming multiple body parts, but the corresponding descriptions are consolidated into a single radiology report. Therefore, matching individual X-rays of the individual body parts to the whole report is *redundant and noisy*, as the report contains irrelevant details from other anatomical regions, depicting a noisy-label problem. To mitigate this, we extract body-part-specific text from the report, enabling the trained VLM to associate each X-ray with its relevant texts. To this end, we first leverage the Large Language Model (LLM) Llama-3.1 [4] to preprocess the text by removing mentions of medical history and irrelevant contextual information. Then, we query the LLM to extract the text corresponding to the available body part XRs of the case (see Fig. 2). The extracted text is then matched to the corresponding XRs, forming *(XR, text, anatomy)* triplets.

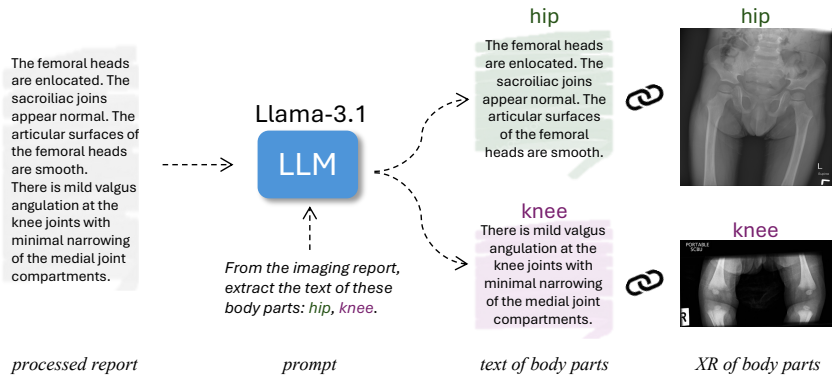

Fig. 2. We leverage Llama-3.1[4] to extract the corresponding text of each body part XR from the raw report as shown in this hip and knee example.

2.2 PedCLIP: Pediatric Vison-Language Model with MoBE

From the curated training set $\mathcal{Q} = \{(x_v, x_t, x_b)_i\}_{i=1}^{N}$, where x_t is the report excerpt of the XR image x_v for the body part index x_b, we train a vision-language model consisting of a text encoder ϕ_t, and an image encoder ϕ_v of L

vision transformer layers to extract their respective features. The text encoder ϕ_t encodes x_t into a sequence of D text tokens $t \in \mathbb{R}^C$ and the global `[CLS]` token: $T = \{t_1, ..., t_D, t_{\texttt{[CLS]}}\}$. The image encoder ϕ_v encodes x_v into a sequence of S patch token $v \in \mathbb{R}^C$ and the global image token `[IMG]`: $V^l = \{v_1, ..., v_S, v_{\texttt{[IMG]}}\}^l$, where C is the feature size and l is the index of the transformer layer.

Existing VLMs [25,2,23,19] only use shared MLP inside the Transformer to learn features. We show empirically that this single-expert design suffers from feature interference when learning from multiple body parts. This phenomenon is also commonly known as negative transfer [14,15] occurring in multi-task learning. Specifically, features from the under-represented body parts are often dominated by those from other body parts. To enhance the anatomical specialization of the image encoder ϕ_v, we adopt a design a Mixture of Body Part Experts (MoBE) module, replacing the MLP to alleviate feature interference.

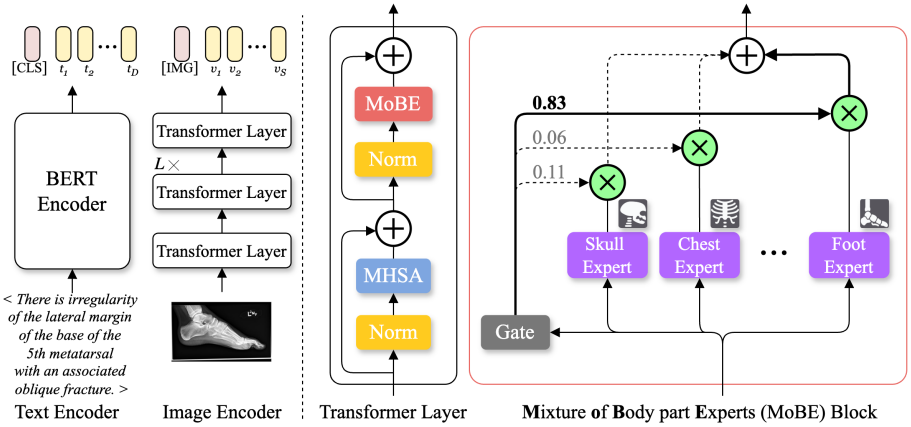

Fig. 3. The architecture of **PedCLIP**: the foundational pediatric vision-language pre-training framework with **M**ixture **o**f **B**ody part **E**xperts (MoBE). *Left*: We adopt BERT Encoder as the text encoder ϕ_t and a Vision transformer of L transformer layers as the image encoder ϕ_v. *Right*: In the Image encoder, we replace the MLP block in the transformer layer with the MoBE block. The MoBE block comprises a gating network G and several MLP experts for each body part. The gating network assigns a weight for each expert output, which is summed to yield the final output of the MoBE block.

Particularly, each MoBE block contains E body part experts, $\{B_e\}^E$ and a gating network G. Each expert B_e is specialized in an anatomical region in the XR images, implemented as an MLP. The gating network G, conditioned on the anatomical region, determines the *involvement* of each expert for a given input. In particular, the MoBE block is constructed as follows. First, the gating network G looks at the global `[IMG]` token of the XR and computes a score for

each anatomical expert B_e such that:

$$s = G(v_{\texttt{[IMG]}}) = \text{softmax}(v_{\texttt{[IMG]}} \cdot W_G), \tag{1}$$

where $s \in \mathbb{R}^E$ is the *expert assignment* scores, indicating which expert B_e is more suitable to process the input image, $W_G \in \mathbb{R}^{C \times E}$ represents the learnable weights of the gating network. Each expert B_e projects the visual tokens such that $B_e(V) = \tilde{V}_e$. Then, the final output of the MoBE block at transformer layer l is computed as the weighted sum across E body experts:

$$\text{MoBE}^l(V^l) = \sum_{e=1}^{E} s^l \tilde{V}_e^l = \sum_{e=1}^{E} \frac{\exp(v^l_{\texttt{[IMG]}} \cdot W_G^l)_e}{\sum^E \exp(v^l_{\texttt{[IMG]}} \cdot W_G^l)} B_e^l(V^l). \tag{2}$$

Thus, the weighting mechanism ensures that the expert corresponding to the relevant anatomical region in the X-ray image predominantly influences the final image token representation. We illustrate the model architecture in Fig. 3.

2.3 Learning objectives

Multi-modal contrastive. As a common practice in VLM pretraining, the training objective is contrastive learning which pulls an XR towards its corresponding report and away from other reports. Here, we employ InfoNCE[17] for multi-modal contrastive learning. Given a minibatch of M XR-reports $\{(x_v, x_t)_i\}^M$, we teach the model to correctly match XR-report pairs in the batch with symmetric cross-modal objectives, including image-to-text and text-to-image losses:

$$\ell_i^{v2t} = -\log\left(\frac{\exp\big(\text{sim}(x_v^i, x_t^i)/\tau\big)}{\sum_{j=1}^{M} \exp\big(\text{sim}(x_v^i, x_t^j)/\tau\big)}\right), \tag{3}$$

$$\ell_i^{t2v} = -\log\left(\frac{\exp\big(\text{sim}(x_t^i, x_v^i)/\tau\big)}{\sum_{j=1}^{M} \exp\big(\text{sim}(x_t^i, x_v^j)/\tau\big)}\right), \tag{4}$$

where $\text{sim}(x_v, x_t) = v_{\texttt{[IMG]}}^{\mathrm{T}} t_{\texttt{[CLS]}}$ denotes the cosine similarity between the global visual and text tokens of the XR and the report, and τ denotes the temperature hyperparameter.

Expert assignment. To correctly assign the input XR to the corresponding body expert, the gating weight W_G of the MoBE is learned with cross-entropy (CE) using the body part label x_b, i.e., $\ell_G = \text{CE}(s, x_b)$. Finally, the overall loss objectives is $\ell = \frac{1}{2}(\ell^{v2t} + \ell^t 2v) + \lambda \ell_G$, where λ scales the expert assignment loss.

3 Experiments and Results

3.1 Experimental settings

Implementation details. PedCLIP uses ViT-B/16 pretrained on ImageNet as the image encoder and BioClinicalBERT [1] as the text encoder. The model is

pretrained on PeXR for 50 epochs with a learning rate of 10^{-5} and a batch size of 256. We employ an image size of 224×224, and pad text reports to a maximum of 120 tokens. In the first 10 epochs, we use a fixed $\lambda = 1$ and teacher forcing, where expert assignments are based on body part labels, x_b, rather than the Gating network's output, for stable early training. After this period, the model begins using the Gating network's output for expert assignments, with λ linearly annealed during the remaining 40 epochs. The MoBE block consists of 23 experts each for one of the 23 anatomical regions. Our models and other baselines were trained on two NVIDIA RTX A6000.

Datasets. We evaluate the model on 5 datasets, encompassing a wide range of body parts. Pedi-CXR [18] is a pediatric Chest XRs dataset that contains 14 pediatric findings. CHD-CXR [28] dataset contains 828 images of congenital heart disease in children. GRAZPEDWRI-DX [16] is a pediatric wrist XRs dataset of 20,037 images for fracture detection. Arthritis is our privately collected multi-body part pediatric XR dataset, which comprises 1,191 images for the pediatric inflammatory arthritis detection task. RSNA-BoneAge [5] is a dataset assessing bone age from 12,800 pediatric hand XRs. The performance of VLMs is evaluated using the AUC, MAE, and Dice scores, respectively, for classification, regression, and segmentation tasks.

Table 1. The zero-shot classification performance (AUC) of the original *adult* pretrained and our reproduced *pediatric* PeXR pretrained vision-language models.

Model	**Venue**	**Dataset**				
		CHD	**PCXR**	**Arth.**	**GRAZ.**	AVG.
Generalist						
BioMedCLIP [27]	NEJM AI'24	0.625	0.513	0.459	0.507	0.526
UniMedCLIP [12]	CoRR'24	0.578	0.482	0.532	0.509	0.525
Adult						
GLoRIA [6]	ICCV'21	0.557	0.538	0.477	0.559	0.533
ChexZero [23]	Nat.BME'22	0.478	0.513	0.475	0.507	0.493
MedCLIP [26]	EMNLP'22	0.521	0.571	0.486	0.494	0.518
MGCA [25]	NIPS'22	0.507	0.468	0.366	0.505	0.461
Prior [2]	ICCV'23	0.505	0.507	0.477	0.491	0.495
CARZero [13]	CVPR'24	0.503	0.511	0.481	0.591	0.522
Pediatric						
GLoRIA [6]	ICCV'21	0.509	0.464	0.491	0.628	0.523
MedCLIP [26]	EMNLP'22	0.511	0.599	0.441	0.513	0.516
MGCA [25]	NIPS'22	0.483	0.529	0.347	0.757	0.529
Prior [2]	ICCV'23	0.467	0.558	0.516	0.507	0.512
CARZero [13]	CVPR'24	0.511	0.596	0.412	0.718	0.559
PedCLIP	**Proposed**	**0.639**	**0.608**	**0.519**	**0.760**	**0.632**

Table 2. The fine-tuning performance of *adult* and *pediatric* VLMs under classification (AUC), regression (MAE) and segmentation (Dice score).

Model	Venue	Classification				Regress.	Segment.
		CHD	PCXR	Arth.	GRAZ.	RSNA	GRAZ.
Adult							
GLoRIA [6]	ICCV'21	0.851	0.742	0.795	0.784	22.2	0.453
MedCLIP [26]	EMNLP'22	0.760	0.774	0.646	0.621	62.3	0.539
MGCA [25]	NIPS'22	0.854	0.763	0.850	0.889	27.9	0.612
Prior [2]	ICCV'23	0.881	0.743	0.813	0.941	14.2	0.494
CARZero [13]	CVPR'24	0.700	0.735	0.594	0.793	21.4	0.641
Pediatric							
GLoRIA [6]	ICCV'21	0.810	0.690	0.854	0.960	16.5	0.569
MedCLIP [26]	EMNLP'22	0.529	0.533	0.623	0.549	12.7	0.479
MGCA [25]	NIPS'22	0.869	0.766	0.877	0.954	21.1	0.619
Prior [2]	ICCV'23	0.892	0.675	0.921	0.945	13.0	0.536
CARZero [13]	CVPR'24	0.812	0.753	0.731	0.935	17.8	0.644
PedCLIP	**Proposed**	**0.911**	**0.782**	**0.934**	**0.970**	**10.6**	**0.652**

Downstream task setup. We evaluate PedCLIP and other foundational vision-language models, including GLoRIA [6], ChexZero [23], MedCLIP [26], MGCA [25], Prior [2] and CARZero [13] on 11 tasks, including both zero-shot and linear-probing classification, segmentation and regression. The performance of the original model pre-trained on the *Adult* data, and pre-trained on our large-scale *Pediatric* PeXR datasets is reported. We also compare with *Generalist* VLMs that are trained on varied body parts and modalities beyond XRs (BioMedCLIP [27], UniMedCLIP [12]). We evaluate the model on zero-shot and linear-probing classification tasks on Pedi-CXR, CHD, Arthritis, and GRAZPEDIWRI-DX, the segmentation task on GRAZPEDIWRI-DX, and the regression task on RSNA-BoneAge. Lastly, we conduct a Concept Alignment task to evaluate the VLM's ability to classify fine-grained concepts.

3.2 Results

Tab. 1 shows **zero-shot classification** performance. The *Adult* pretrained models struggle with pediatric data, demonstrating comparable results with the *Generalist* models. Adult models are better on CHD as they are chest-pretrained, and mixed body parts in PeXR interfere with chest features of PeXR-VLMs (MoBE fixes this in Fig.5). Notably, retraining the models on our pediatric dataset PeXR yields marked improvements across the board. This necessitates dedicated pediatric foundational models. Our PedCLIP significantly improves the score by 10.6%, 9.9%, and 7.2% over the best-performing models in *Generalist*, *Adult*, and *Pediatric* settings respectively . This highlights the effectiveness of our MoBE for multi-anatomy pre-training. Tab. 2 shows the fine-tuning evaluation of classification, regression, and segmentation performance. The proposed PedCLIP achieves

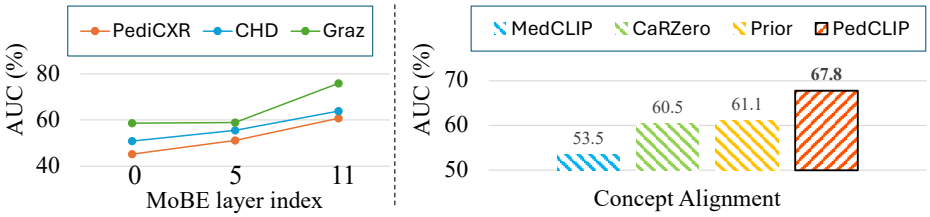

Fig. 4. *Left*: MoBE's performance as integrated in different vision transformer layer. *Right*: Concept Alignment performance of MoBE versus other VLMs.

a 5% and 3% AUC improvement over the *Adult* and *Pediatric* pretrained VLMs, respectively. To evaluate the **alignment of models with pediatric inflammatory arthritis findings**, six clinical concepts are annotated on our Arthritis dataset: *inflammation, bony changes, erosion, joint space narrowing, soft tissue swelling, and effusion*. We then perform a linear probing evaluation for each concept and report the results in Fig. 4 - *Right*. PedCLIP achieves the highest performance, suggesting a strong alignment with human-interpretable concepts.

3.3 Ablation Study

MoBE layer index. We investigate the optimal layer for integrating the MoBE block, as shown in Fig. 4.*Left*, indicating that MoBE performs best when placed in the later layers of the image encoder.

MoBE design. Using a single MLP in the transformer layer of ϕ_v on a multi-body part dataset like PeXR yields suboptimal results due to *feature interference*, where features of one part are forgotten when the model learns other parts.

As shown in Fig. 5, the single MLP performs better when only trained on the target body part, but the performance drops when trained on multiple body parts. In contrast, MoBE handles multi-body part datasets effectively, matching the MLP-target settings. The learned soft-gate enables shared findings features across body parts. While there are other MoE approaches for medical domain [10,3], to our knowledge, this is the first to apply soft-MoE to anatomically diverse pediatric data.

4 Conclusion

In conclusion, this paper demonstrates the limitations of current medical VLMs on pediatric medical tasks and proposes PedCLIP, the first pediatric VLM, trained on PeXR, a large-scale pediatric XR multi-body part dataset, achieving state-of-the-art performance. By incorporating MoBE, where each expert specializes in a particular body part, PedCLIP effectively addresses the heterogeneity challenge of the multi-body part nature of the pediatric imaging data. We believe this work will advance foundational models in pediatric imaging analysis.

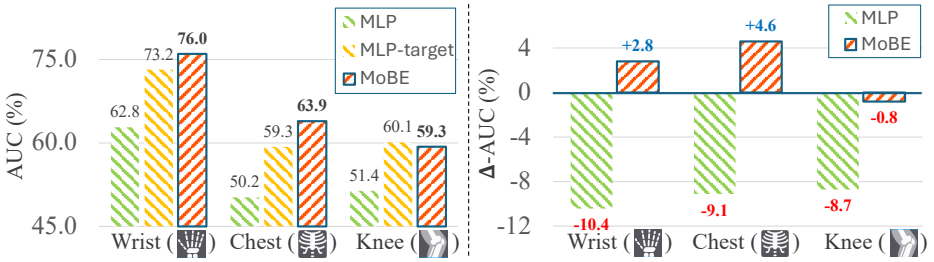

Fig. 5. *Left*: Zero-shot AUC on Wrist (GRAZ.), Chest (CHD-CXR) and Knee (Arthtitis) dataset of 3 settings: MLP (trained on multiple body-part PeXR), MLP-target (trained only on the target body part samples in PeXR) and MoBE (trained on PeXR). *Right*: the relative AUC compared to the MLP-target settings of multi-MLP and MoBE, which indicates *feature interference*.

Acknowledgments. This study was funded by the Hospital Research Foundation Group, whose support made this work possible and is gratefully acknowledged.

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

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
2. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X.: Prior: Prototype representation joint learning from medical images and reports. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21361–21371 (2023)
3. Chopra, S., Mao, L., Sanchez-Rodriguez, G., Feola, A.J., Li, J., Kira, Z.: Medmoe: Modality-specialized mixture of experts for medical vision-language understanding. arXiv preprint arXiv:2506.08356 (2025)
4. Dubey, A., Jauhri, A., Pandey, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
5. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., et al.: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019)
6. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
7. Huy, T.D., Huyen, H.C., Nguyen, C.D., et al.: Adversarial contrastive fourier domain adaptation for polyp segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
8. Huy, T.D., Huynh, D.A., Xie, Y., Qi, Y., Chen, Q., Nguyen, P.L., Tran, S.K., Phung, S.L., Hengel, A.v.d., Liao, Z., et al.: Seeing the trees for the forest: Rethinking weakly-supervised medical visual grounding. arXiv preprint arXiv:2505.15123 (2025)

9. Huy, T.D., Tran, S.K., Nguyen, P., Tran, N.H., Sam, T.B., van den Hengel, A., Liao, Z., Verjans, J.W., To, M.S., Phan, V.M.H.: Interactive medical image analysis with concept-based similarity reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30797–30806 (2025)
10. Jain, Y., Behl, H., Kira, Z., Vineet, V.: Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. Advances in Neural Information Processing Systems **36**, 69625–69637 (2023)
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. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. arXiv preprint arXiv:2412.10372 (2024)
13. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11137–11146 (2024)
14. Li, D., Nguyen, H., Zhang, H.R.: Identification of negative transfers in multitask learning using surrogate models. Transactions on Machine Learning Research (2023), featured Certification
15. Liu, S., Liang, Y., Gitter, A.: Loss-balanced task weighting to reduce negative transfer in multi-task learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9977–9978 (2019)
16. Nagy, E., Janisch, M., Hržić, F., Sorantin, E., Tschauner, S.: A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. Scientific data **9**(1), 222 (2022)
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Pham, H.H., Nguyen, N.H., Tran, T.T., Nguyen, T.N., Nguyen, H.Q.: Pedicxr: an open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. Scientific Data **10**(1), 240 (2023)
19. Phan, V.M.H., Xie, Y., Qi, Y., Liu, L., Liu, L., Zhang, B., Liao, Z., Wu, Q., To, M.S., Verjans, J.W.: Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11492–11501 (2024)
20. Quan, T.M., Thanh, H.M., et al.: Xpgan: X-ray projected generative adversarial network for improving covid-19 image classification. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1509–1513. IEEE (2021)
21. Rajpurkar, P., Irvin, J., Zhu, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
23. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature biomedical engineering **6**(12), 1399–1406 (2022)
24. U.S. Food and Drug Administration: Pediatric x-ray imaging (2025), https://www.fda.gov/radiation-emitting-products/medical-imaging/pediatric-x-ray-imaging, accessed: 2025-02-12

25. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning (supplementary material). In: Advances in Neural Information Processing Systems (2022)
26. 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. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
27. Zhang, S., Xu, Y., Usuyama, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
28. Zhixin, L., Gang, L., Zhixian, J., Sibao, W., Silin, P.: Chd-cxr: a de-identified publicly available dataset of chest x-ray for congenital heart disease. Frontiers in Cardiovascular Medicine **11**, 1351965 (2024)