CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification

Supplementary Material

A. Comparison with State-of-the-art Methods

In this section, we present a detailed comparison of our proposed CARZero with the existing state-of-the-art (SOTA) image-text pre-training methods in various datasets under zero-shot classification settings.

Figure 1, Table 1, Table 2, and Table 3 present the comprehensive results on the Open-I [2], ChestXray14 [8] CheXpert [4], and ChestXDet10 [5] datasets, respectively. These results demonstrate the efficacy of our method in discerning distinct disease categories, showcasing our approach's robust performance in comparison to existing methodologies.

Figure 3, Figure 4, and Figure 5 showcase the results on the PadChest [1] dataset. These figures offer a detailed insight into the performance of our method, emphasizing its effective handling of the dataset with long-tailed multi-label classification challenge. Our proposed CARZero method outperforms existing SOTA methods in both the head and tail classes, which demonstrates its effectiveness and generalizability.

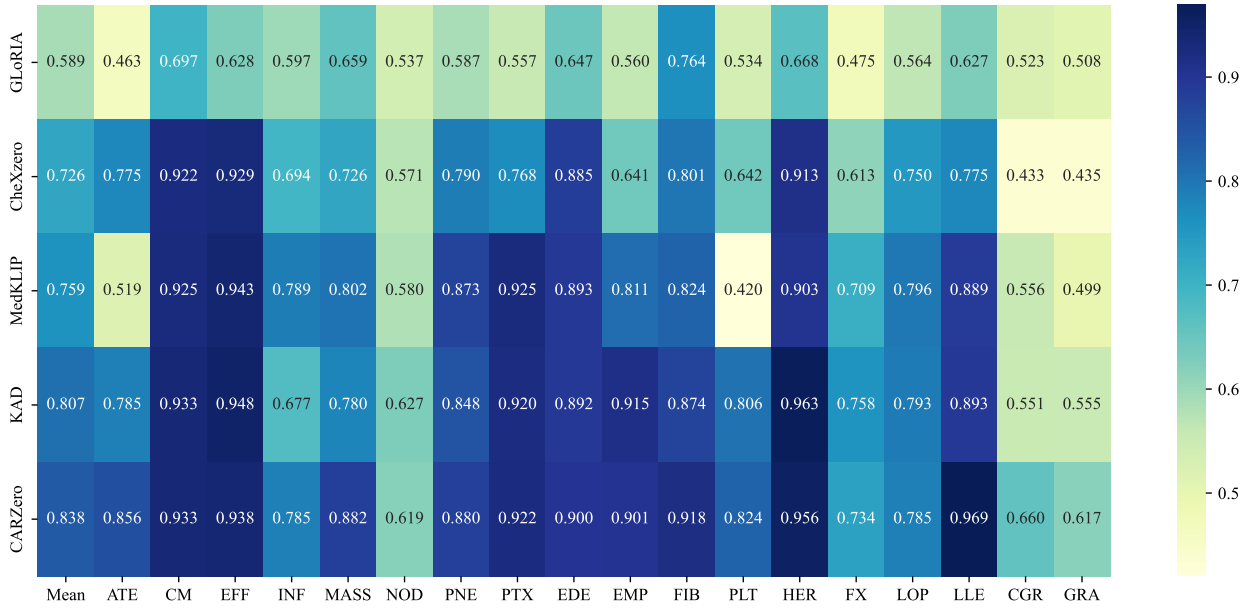


Figure 1. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to Open-I dataset across 18 disease categories in terms of AUC performance. The abbreviations ATE, CM, EFF, INF, MASS, NOD, PNE, PTX, EDE, EMP, FIB, PLT, HER, FX, LOP, LLE, CGR, and GRA correspond to Atelectases, Cardiomegaly, Effusion, Infiltrate, Mass, Nodule, Pneumonia, Pneumothorax, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, Fracture, Lung Opacity, Lung Lesion, Calcified Granuloma, and Granuloma, respectively.

Method	Mean	ATE	CM	EFF	INF	MASS	NOD	PNE	PTX	CON	EDE	EMP	FIB	PLT	HER
GLoRIA [3]	0.610	0.653	0.704	0.762	0.660	0.613	0.508	0.587	0.572	0.697	0.762	0.499	0.459	0.613	0.450
CheXzero [6]	0.712	0.705	0.837	0.850	0.652	0.706	0.598	0.717	0.805	0.773	0.853	0.430	0.655	0.555	0.824
MedKLIP [9]	0.726	0.671	0.842	0.813	**0.706**	0.742	0.621	0.698	0.821	0.719	0.803	0.783	0.604	0.499	0.841
KAD [10]	0.789	0.770	**0.854**	0.824	0.694	0.754	0.698	0.734	0.860	0.718	0.809	**0.879**	0.780	0.718	**0.952**
CARZero	**0.811**	**0.819**	0.852	**0.873**	0.670	**0.854**	**0.718**	**0.737**	**0.871**	**0.786**	**0.884**	0.808	**0.788**	**0.770**	0.928

Table 1. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to ChestXray14 dataset across 14 disease categories in terms of AUC performance. The abbreviations ATE, CM, EFF, INF, MASS, NOD, PNE, PTX, CON, EDE, EMP, FIB, PLT, and HER correspond to Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia, respectively.

Method	Mean	Aelectasis	Cardiomegaly	Consolidation	Edema	Pleural effusion
GLoRIA [3]	0.750	0.807	0.802	0.588	0.747	0.807
CheXzero [6]	0.889	0.816	0.906	0.892	0.897	0.932
MedKLIP [9]	0.879	0.813	0.866	0.858	0.911	0.947
KAD [10]	0.905	**0.884**	0.885	0.865	0.943	**0.949**
CARZero	**0.923**	0.879	**0.916**	**0.923**	**0.950**	**0.949**

Table 2. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to CheXpert dataset across 5 disease categories in terms of AUC performance.

Method	Mean	ATE	CALC	CONS	EFF	EMPH	FIB	FX	MASS	NOD	PTX
GLoRIA [3]	0.645	0.622	0.524	0.718	0.866	0.607	0.523	0.494	0.725	0.630	0.744
CheXzero [6]	0.640	0.622	0.392	0.826	0.873	0.503	0.628	0.665	0.650	0.479	0.766
MedKLIP [9]	0.713	0.746	0.527	0.831	0.905	0.728	0.567	0.642	0.796	0.572	0.814
KAD [10]	0.735	0.757	0.563	0.824	0.888	0.888	0.687	0.608	0.695	0.566	0.874
CARZero	**0.796**	**0.782**	**0.642**	**0.857**	**0.910**	**0.926**	**0.736**	**0.714**	**0.843**	**0.637**	**0.915**

Table 3. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to ChestXDet10 dataset across 10 disease categories in terms of AUC performance. The abbreviations ATE, CALC, CONS, EFF, EMPH, FIB, FX, MASS, NOD, and PTX correspond to Atelectasis, Calcification, Consolidation, Effusion, Emphysema, Fibrosis, Fracture, Mass, Nodule, and Pneumothorax, respectively.

Prompting LLM to generate prompt template

messages = [{"role": "user", "content": " You now assume the role of a knowledgeable radiologist. Please assist me in reading medical reports, extracting disease information according to given diseases name, and generating this information using a fixed template. The template format includes: "There is [D]", "There may be [D]", and "There is no [D]", where [D] represents the describing information of diseases obtained from the report. Please generate sentences according to the given format one by one.
Given an Example:
{"role": "user", "content": "Report: as compared to the lung volumes have slightly decreased. signs of mild over inflation and moderate pleural effusion persist. moderate cardiomegaly . elongation of the descending aorta. no pneumonia . Disease: effusion, cardiomegaly, pneumonia. " ,
"role": "system", "content": "There is moderate pleural effusion. There is moderate cardiomegaly. There is no pneumonia."}
Example END."]
messages.append({"role": "user", "content": Report: {**report**}
Disease: {**keywords**}**)**

Figure 2. Messages are used to prompt LLM for generating prompt templates from medical reports and diseases. Diseases are derived through keyword matching from reports related to 14 common diseases in the ChestXray14. Each prompt includes a manually curated one-shot example for enhanced specificity and relevance.

We also compare ViLLA [7], which introduces a self-supervised multimodal representation learning approach designed to capture fine-grained region-attribute relationships from complex datasets. Since ViLLA doesn't release its model, we compare the accuracy on CheXpert5x200, where CARZero surpasses ViLLA (56.5 vs. 55.9, reported in [7]).

B. Prompt Alignment

As shown in Figure 2, prompting instruction is utilized to generate the prompt template within the MIMIC reports.

C. Input of Text encoder

CARZero utilizes an LLM to align reports following a standardized template. We use both original ('ORI') and generated ('GEN') reports for training. This approach not only preserves relevant diagnostic information but also aligns the prompts. In the training stage, each report is divided into multiple sentences. Following the official implementation in GLoRIA [3], in each iteration, we randomly select one sentence for training, which can expand the diversity of image-text pairs.

ORI	GEN	AUC	MCC	F1	ACC	AUPRC
✓		0.801	0.241	0.253	0.821	0.220
	✓	0.796	0.242	0.255	0.846	0.213
✓	✓	**0.810**	**0.257**	**0.270**	**0.867**	**0.224**

D. Computational Efficiency

In terms of computational efficiency, CARZero significantly surpasses KAD regarding throughput and floating-point operations per second (flops), which underscores its clinical relevance.

Model	Params ↓	Flops ↓	Throughput ↑
KAD [10]	**111M**	35.41G	203 img/s
CARZero	179M	**25.12G**	**413 img/s**

E. Additional Visualization Results

In this section, we delve deeper into the visualization aspects of our study, focusing on the comparative and analytical visual representations that further substantiate the effectiveness of our proposed method. These additional visualizations not only provide empirical evidence supporting our model's superiority in various scenarios but also offer intuitive insights into the intricate workings of the model. By presenting these results, we aim to offer a more comprehensive understanding of how our approach advances the field of zero-shot learning in medical image analysis.

Figure 6 and Figure 7 provide a comparative visualization of the attention maps between our proposed CARZero and MedKLIP [9] in zero-shot setting. These figures demonstrate how CARZero more precisely identifies the specific locations of diseases in CXR images. The accuracy in pinpointing disease-specific regions highlights CARZero's advanced capabilities in medical image analysis. Notably, our method shows a significant advantage in detecting smaller lesions, such as Calcification, Fracture, and Nodules, with attention maps that accurately locate these areas. This demonstrates the high interpretability of our method.

Figure 8 showcases the t-SNE visualization of Similarity Representations (SimR) of images and texts within CARZero. Figure 8a illustrates distinct clustering of SimRs within identical categories, while ensuring clear demarcation among different categories. This observation confirms the effectiveness of our proposed SimR in capturing the complex relationships between medical images and corresponding textual descriptions, thus aligning the modalities with high precision. Moreover, the illustrations from Figure 8b to Figure 8f show that the SimR for both positive and negative cases is distinctly clustered for various diseases, indicating that our derived SimR contains valid semantic information at the disease level. This outcome highlights the precision and semantic accuracy of our method in the domain of radiology zero-shot classification, thereby enhancing its applicability and utility in clinical practice.

Figure 3. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to the human-annotated subset of the PadChest dataset (totalling 39,053 chest X-rays and 192 classes). The results for the **1-64** classes are shown here. Mean AUC are shown for each class, and n refers to the number of positive samples.

Figure 4. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to the human-annotated subset of the PadChest dataset (totalling 39,053 chest X-rays and 192 classes). The results for the **65-128** classes are shown here. Mean AUC are shown for each class, and n refers to the number of positive samples.

Figure 5. Comparative analysis between the existing zero-shot classification approaches and our proposed CARZero method, applied to the human-annotated subset of the PadChest dataset (totalling 39,053 chest X-rays and 192 classes). The results for the **129-192** classes are shown here. Mean AUC are shown for each class, and n refers to the number of positive samples.

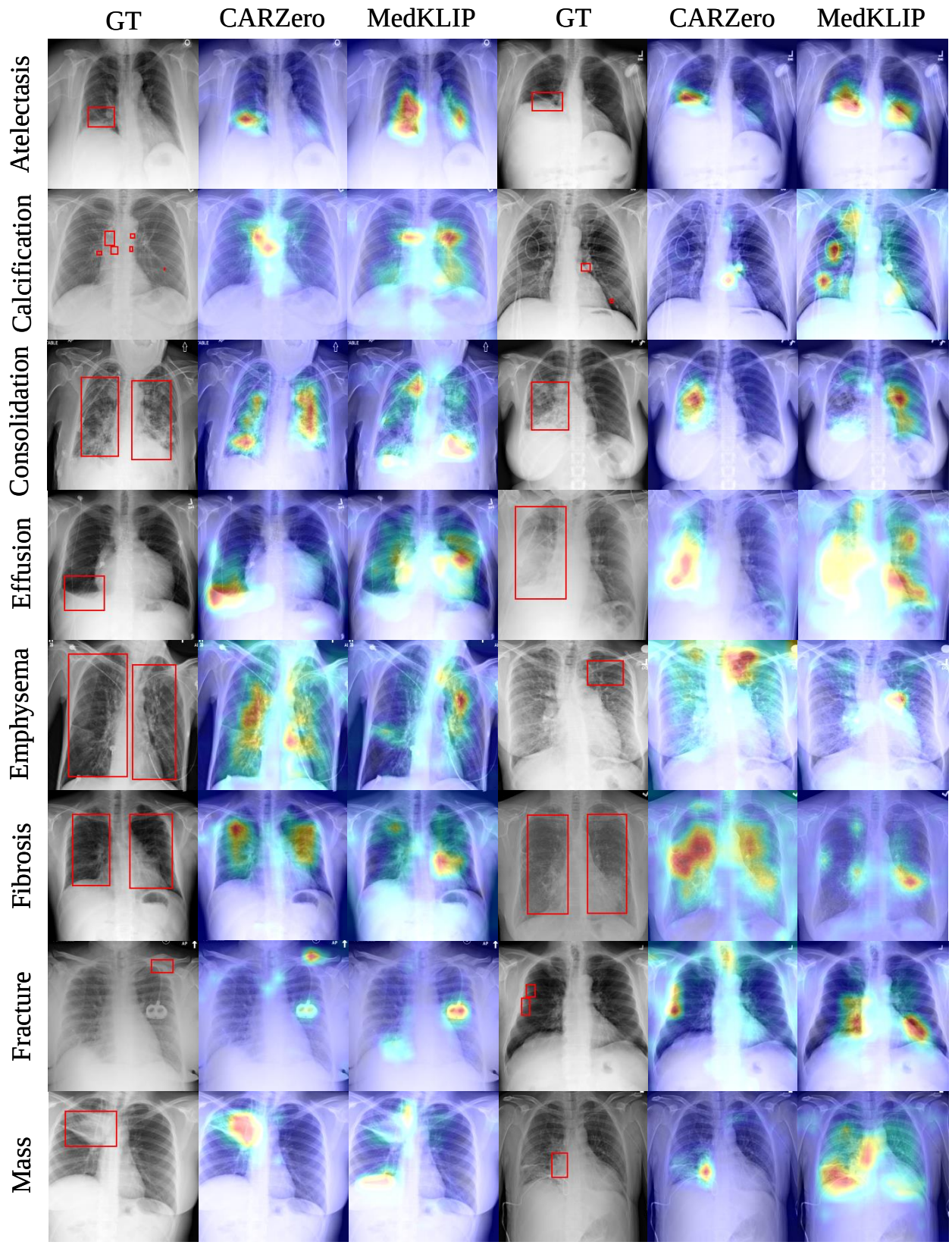


Figure 6. Comparative visualization of attention maps between CARZero and MedKLIP on ChestXDet10 in zero-shot setting. This visualization displays the attention maps for classes **1-8**, where red boxes highlight the ground truth areas for detection. Areas with higher activation weights, indicating stronger correlations between specific words and image regions, are represented by highlighted pixels.

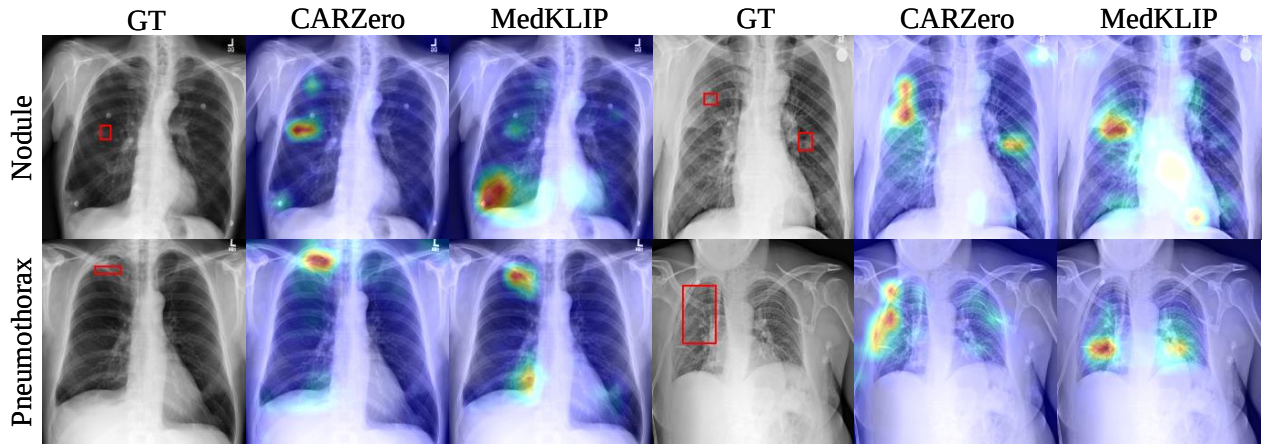
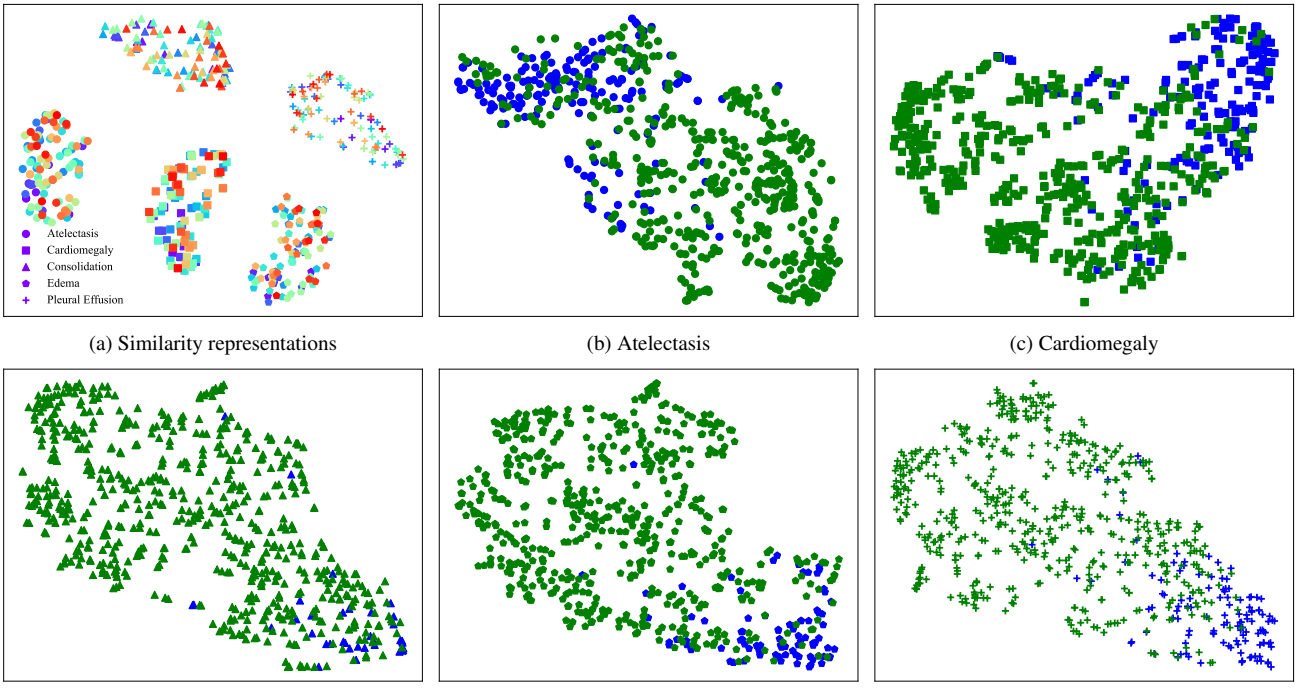


Figure 7. Comparative visualization of attention maps between CARZero and MedKLIP on ChestXDet10 in zero-shot setting. This visualization displays the attention maps for classes **9-10**, where red boxes highlight the ground truth areas for detection. Areas with higher activation weights, indicating stronger correlations between specific words and image regions, are represented by highlighted pixels.



(a) Similarity representations (b) Atelectasis (c) Cardiomegaly

(d) Consolidation (e) Edema (f) Pleural Effusion

Figure 8. t-SNE visualization of similarity representations between images and texts for 5 classes in CheXpert. (a) Different colors represent various CXR images, while distinct shapes indicate different diseases. A subset of 100 CXR images is randomly selected for clearer visualization. (b)-(f) display the similarity representations within each class, where blue signifies positive cases and green indicates negative cases. The total testing set of CheXpert is represented here.

References

[1] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 1

[2] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 1

[3] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 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*, pages 3942–3951, 2021. 1, 2

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 1

[5] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*, 2020. 1

[6] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 1, 2

[7] Maya Varma, Jean-Benoit Delbrouck, Sarah Hooper, Akshay Chaudhari, and Curtis Langlotz. Villa: Fine-grained vision-language representation learning from real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22225–22235, 2023. 2

[8] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 1

[9] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01, 2023. 1, 2, 3

[10] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023. 1, 2, 3