Supplementary for "PairAug: What Can Augmented Image-Text Pairs Do for Radiology?"

Yutong Xie[1*] Qi Chen[1*] Sinuo Wang[1] Minh-Son To[4] Iris Lee[4] Ee Win Khoo[4]
Kerolos Hendy[4] Daniel Koh[4] Yong Xia[2,3] Qi Wu[1†]

[1] *Australian Institute for Machine Learning (AIML), The University of Adelaide, Australia*
[2] *School of Computer Science and Engineering, Northwestern Polytechnical University, China*
[3] *Ningbo Institute of Northwestern Polytechnical University, China*
[4] *South Australia Medical Imaging, Australia*

yutong.xie678@gmail.com, {qi.chen04, qi.wu01}@adelaide.edu.au

Appendix

This document provides more discussions and experimental details to supplement the main submission. We organise the supplementary into the following sections.

- In section A, we depict more implementation details.
- In section B, we show more examples of newly generated reports by large language model.
- In section C, we provide some visual results of failure cases.
- In section D, we provide a discussion on the impact of hyper-parameters.

A. More Details of Experimental Setup

A.1. Computational Cost

In our method, different parts have varying computational requirements, as indicated in Table A. For the Pairwise Augmentation (PairAug) part, we use eight NVIDIA A100 GPUs, costing a total time of 24 hours. This high computational demand can be attributed to the inference complexity of diffusion models. The medical visual-language pre-training (MedVLP) is trained on a less resource-intensive NVIDIA 3090 GPU, costing a total time of 5 hours. The downstream tasks are also performed on an NVIDIA 3090 GPU, where the fine-tuning task takes no more than one hour for training and less than 1 second for testing. Despite the high computational expenditure of the PairAug and MedVLP parts, the fast downstream training and online testing processes indicate the feasibility of incorporating our approach into regular clinical workflows.

A.2. Experimental Details for Downstream

We test the performance of learned pre-trained representations on three radiology-based downstream datasets: (1) For CheXpert dataset [3], we use its official test set for zero-shot evaluation aiming to classify each image into 5 five individual binary labels: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion; (2) For PadChest dataset [1] e adopt 39,053 chest X-rays annotated by board-certified radiologists for zero-shot evaluation. It has 193 disease image labels, including 174 radiographic findings and 19 differential diagnoses; and (3) For the RSNA Pneumonia dataset [5], we seek to classify each radiograph as negative or positive for pneumothorax. We divided the RSNA dataset into training, validation, and test sets with a ratio of 80%/10%/10%, respectively. The input size is set to 224×224. We also employ online data argumentation to enlarge the training dataset. We optimise the downstream network with the AdamW [4] algorithm with cross-entropy loss and empirically set the initial learning rate to 0.0005, the batch size to 96, and the max epochs to 50.

For the zero-shot evaluation, we adopt a methodology inspired by the work [6]. This involves using labels from the test set to generate both positive and negative prompts for each condition. Specifically, for a given label, we create a positive prompt, such as '<label>', and a corresponding negative prompt, 'no <label>', to facilitate the softmax evaluation process. The evaluation procedure is structured as follows: Initially, we calculate logits – a type of raw output from the last layer of neural networks before applying the softmax function – using both positive and negative prompts. For instance, for the label 'atelectasis', we compute logits

Table A. Computational Cost.

	PairAug	MedVLP	Downstream
GPUs	eight A100	one 3090	one 3090
Time Cost	24 hours	5 hours	one hour

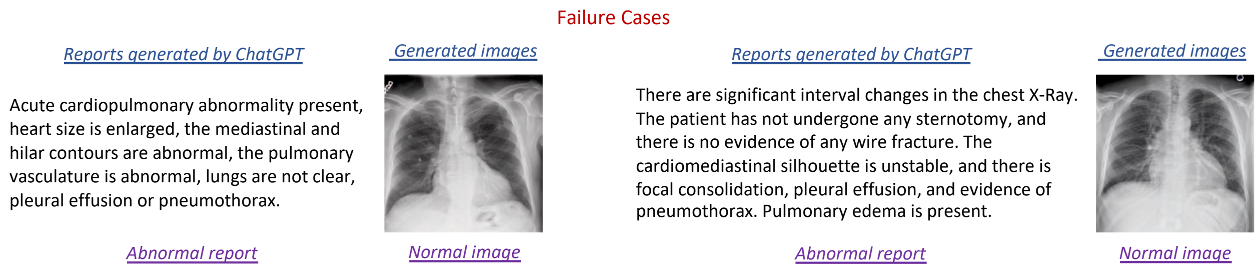

Figure A. Visual Results of Failure Cases

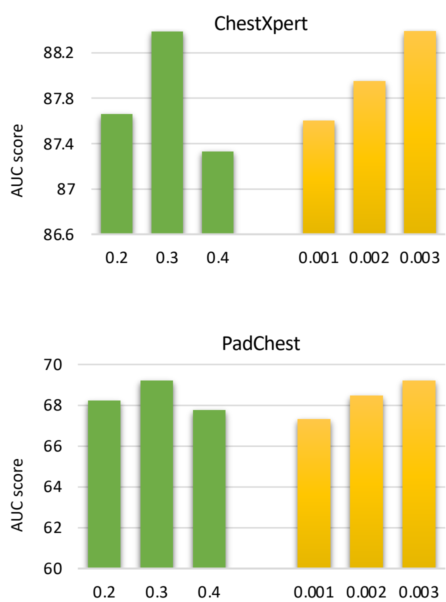

Figure B. AUC scores on ChestXpert and PadChest datasets vs. hyper-parameters values on thresholds τ (green bars) and ϵ (yellow bars).

for 'atelectasis' as the positive prompt and 'no atelectasis' as the negative prompt. Following this, we apply the softmax function to these logits, creating a probabilistic comparison between the positive and negative scenarios. Finally, the softmax probabilities derived from the positive logits are interpreted as the likelihood of the corresponding disease being present in the chest X-ray image.

B. Samples of New Reports Generated by LLM

We provide more samples of new radiology reports generated by ChatGPT. As shown in Figure C, given the prompt of "*Following is an original chest X-Ray report. Generate one possible augmentation that is limited to 50 words while conveying partial opposite meanings than the original report*" with original chest X-Ray reports, ChatGPT exhibits the ability to produce varied reports. This ability to simulate the creation of new clinical scenarios contributes significantly to generating novel information continually. Nevertheless, we must also acknowledge the limitations of the current report generation strategy. As highlighted in the primary manuscript, ChatGPT occasionally produces informal terminology. This observation underscores potential avenues for future enhancement, such as creating more sophisticated prompts or implementing filtering mechanisms to expunge these casual terms. Moreover, the current strategy for generating reports relies on including original reports as inputs, which may constrain the diversity of the newly synthesised information. This limitation underscores the necessity to devise more diverse prompts or to develop a strategy that is not reliant on original reports, thereby enhancing the variety and richness of the augmented data.

C. Some Visual Results of Failure Cases

In our study, some failure cases exhibited a clear disconnect between the semantic content of the generated report and the corresponding generated image, as shown in Figure A. These instances of misalignment can potentially be attributed to the limitations of the employed generation model, specifically, the RoentGen model [2]. The RoentGen model's training regimen primarily focused on the 'impression' sections of the radiology reports. While this approach allowed it to develop a general understanding of the diagnostic conclusions drawn from the images, it might not have adequately exposed the model to the intricate details contained within the 'findings' sections. These sections often contain detailed descriptions of specific abnormalities present in the images, and the omission of this data in training might limit the model's ability to generate fully accurate and semantically aligned text and images. Furthermore, the model's performance may also be affected by the imbalance in the representation of normal and abnormal findings in the training data. Particularly, the model might struggle to accurately represent images with multiple or rare abnormalities due to insufficient exposure during training. These limitations also inspire further refinements in the generation model, including more detailed findings in the training stage, which might help improve the semantic alignment between the generated reports and images.

D. Impact of Thresholds τ and ϵ

In Figure B, we investigate the influence of the thresholds τ and ϵ in shaping the model's performance on the downstream tasks. Analysing the impact of τ, we notice an increase in performance when we move the threshold from 0.2 to 0.4 on both the ChestXpert and PadChest datasets. This suggests that a higher threshold for τ allows us to filter out low-quality image-text pairs generated by the InterAug branch, thus enhancing the overall quality of the training data. However, further increasing τ to 0.4 results in a slight performance drop. This indicates that when the threshold becomes too high, the data pruning might become too aggressive, excluding potentially beneficial image-text pairs and causing a slight performance decrease. Turning our attention to ϵ, we observe a substantial performance increase on both datasets when the threshold increases from 0.001 to 0.003. This improvement suggests that a non-zero threshold for ϵ can help selectively incorporate the higher-quality synthesised data from the IntraAug branch. These results illustrate the importance of carefully setting the thresholds τ and ϵ for the model to use the synthesised data effectively. The optimal threshold settings would strike a balance: they would allow for the inclusion of beneficial synthesised data while filtering out potentially detrimental low-quality pairs, thus leading to improved performance on downstream tasks.

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] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 2

[3] 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

[4] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 1

[5] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019. 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

Prompt: *Following is an original chest X-Ray report. Generate one possible augmentation that is limited to 50 words while conveying partial opposite meanings than the original report*

Real-world reports	***New reports generated by ChatGPT***
Comparison is made to previous study from ___,there is a dobbhoff tube whose distal tip is in the body of the stomach, there are bilateral pleural effusions, there is a right sided pleural based catheter, there is no pneumothoraxes or signs for overt pulmonary edema, overall these findings are stable since prior study from ___.	*Comparison is made to the previous study from ___, there is no dobbhoff tube, bilateral pleural effusions are resolved, there is no right-sided pleural-based catheter, there are signs of pneumothoraxes and overt pulmonary edema. Overall, these findings are unstable compared to the prior study from ___.*
Mild thoracic scoliosis with subsequent asymmetry of the rib cage, the lung volumes are normal, normal appearance of the cardiac silhouette, no pleural effusions no pneumothorax, no lung parenchymal abnormalities such as pneumonia or pulmonary edema, normal size of the cardiac silhouette, normal hilar and mediastinal structures.	*Moderate thoracic scoliosis with significant asymmetry of the rib cage, the lung volumes are reduced, abnormal appearance of the cardiac silhouette, presence of pleural effusions and pneumothorax, lung parenchymal abnormalities such as pneumonia and pulmonary edema, enlarged size of the cardiac silhouette, abnormal hilar and mediastinal structures.*
Mild left basal atelectasis, otherwise unremarkable, no definite displaced rib fracture though if there is continued concern dedicated rib series may be performed to further assess, ap upright and lateral views the chest were provided, mild left basal atelectasis, lungs are otherwise clear, no signs of pneumonia or edema, no large effusion or pneumothorax, cardiomediastinal silhouette is normal, bony structures are intact, no free air below the right hemidiaphragm.	*Moderate left basal hyperinflation, otherwise notable, possible displaced rib fracture, a dedicated rib series should be performed to further evaluate. An upright and lateral view of the chest were given. Moderate left basal hyperinflation, lungs show signs of pneumonia and edema, large effusion or pneumothorax present. Cardiomediastinal silhouette is abnormal, bony structures are compromised, free air is present below the right hemidiaphragm.*
No evidence of pneumonia edema or effusion, a left sided picc is unchanged in position, cardiac and mediastinal contours are unchanged from the prior exam, there is no evidence of pulmonary edema, no effusions are identified, there is no pneumothorax, surgical chain sutures are again seen in the right upper lobe consistent with prior surgery, again fullness to the right suprahilar region likely corresponds to a consolidative fibrotic area seen on recent chest ct dated ___.	*There is evidence of pneumonia edema and effusion, a left-sided picc is misplaced, cardiac and mediastinal contours are altered from the prior exam, there is evidence of pulmonary edema, effusions are identified, there is a pneumothorax, surgical chain sutures are not seen in the right upper lobe inconsistent with prior surgery, again fullness to the right suprahilar region likely corresponds to a resolving fibrotic area seen on recent chest ct dated ___.*
AP chest compared to ___ multifocal pulmonary consolidation is clearing, heart is normal size, right paratracheal mediastinal fullness due in part to adenopathy is longstanding, no pleural effusion, heart size normal, right jugular line ends in the region of the superior cavoatrial junction.	*AP chest compared to ___ multifocal pulmonary consolidation is worsening, heart is enlarged, right paratracheal mediastinal fullness due in part to tumor is new, pleural effusion present, heart size increased, right jugular line extends beyond the superior cavoatrial junction.*
As compared to the previous radiograph the lung volumes have slightly decreased, there is minimal fluid overload in both the vascular and interstitial compartment, normal size of the cardiac silhouette, moderate tortuosity of the thoracic aorta, no pleural effusions, no pneumonia.	*Compared to the previous radiograph, the lung volumes have slightly increased, there is significant fluid overload in both the vascular and interstitial compartment, enlarged cardiac silhouette, severe tortuosity of the thoracic aorta, presence of pleural effusions and pneumonia.*
Clear lungs, the lung volumes are low, the lungs are clear without pleural effusion or pneumothorax, the aorta is unfolded, the heart size is normal.	*The lungs are congested with high volumes, the lungs are not clear with pleural effusion or pneumothorax, the aorta is folded, the heart size is enlarged.*
Unchanged mild to moderate cardiomegaly without a superimposed acute intrathoracic process, the lung volumes are hyperinflated and the lungs are clear, there is no pneumothorax or focal airspace consolidation, the heart is mild to moderately enlarged but unchanged from at least ___,there is no evidence for pulmonary edema, slight blunting of the costophrenic angles may reflect trace pleural effusions unchanged, the mediastinal hilar structures are unremarkable.	*Significant cardiomegaly with an acute intrathoracic process, the lung volumes are reduced and the lungs are hazy. There is evidence of pneumothorax and focal airspace consolidation. The heart is severely enlarged and worsening from at least ___. There is evidence of pulmonary edema, and marked blunting of the costophrenic angles may reflect significant pleural effusions. The mediastinal hilar structures are abnormal.*
In comparison with study of ___ the endotracheal tube and swan ganz catheter have been removed, the left chest tube remains in place and there is no evidence of pneumothorax, mild atelectatic changes are seen at the left base.	*Unlike the previous examination, the endotracheal tube and swan ganz catheter have been inserted, the left chest tube has been removed, and there is evidence of pneumothorax. Significant atelectatic changes are observed at the left base.*
heart size is normal, this ascending aorta is unchanged, lungs are essentially clear except for chronic interstitial changes bilaterally, the findings might be attributed to the provided history of copd, if clinically warranted correlation with cross sectional imaging might be considered, no pleural effusion or pneumothorax is seen.	*Heart size is abnormal, the ascending aorta is enlarged, lungs are significantly congested with acute interstitial changes bilaterally. The findings might be attributed to the provided history of pneumonia. If clinically warranted, correlation with cross-sectional imaging might be considered. Pleural effusion or pneumothorax is seen.*

Figure C. More samples of new reports generated by ChatGPT. '___' means anonymous processing.