Data We use images and pneumonia annotations from several datasets for training, including COVID_QU_EX_all [1], covid_rural [2], mastermind [3], RSNA pneumonia detection challenge [4], and QaTA_COV19 [5]. Pneumonia localization masks are available in the RSNA pneumonia detection challenge, QaTA_COV19, COVID_QU_EX_all, and covid_rural datasets, and are used to refine the attention maps. The number of infection masks in these datasets ranges from 221 to 2,913. We use the training splits from all datasets to optimize the model's classification loss. Additionally, the validation splits from datasets with pneumonia opacity masks are included in the training process to refine attention maps. Model comparisons are conducted locally using test sets provided by the challenges and in-house annotated images from the mastermind dataset.

Model Architecture We use a pre-trained, frozen text encoder and an image encoder within a CLIP framework [6], with the text encoder prompted by the word "pneumonia." The text encoder is initialized with weights from Med-KEBERT [7], while the image encoder is a Vision Transformer (ViT) that converts input images into embeddings. A trainable cross-attention layer integrates the embeddings from both the text and image encoders. The initial weights for the image encoder and cross-attention layer are inherited from DeViDe [8]. To refine the attention map and better align it with ground truth masks, we employ a disease-specific attention refinement model. We utilize student and teacher model for the class and heatmap prediction. The student model is updated using back-propagation while the teacher model's gradients are stopped and the weights are shared from student. For the much higher penalty of the false-negative pixels than the false-positive pixels, we update the teacher model for less epochs and predict the heatmap for looser active region, and the student model will be trained for more epochs for more precise class prediction. Our approach incorporates three types of losses: image-level cross-entropy loss, dice loss, and pixel-level cross-entropy loss.

Training Scheme For images with pneumonia localization bounding boxes, we compute the pixel-level loss, while classification loss is applied to all images. The model is optimized using the AdamW optimizer, with an initial learning rate of 5×10^{-5}, and a CosineLRScheduler for learning rate adjustment. The training is conducted on V100 16G GPUs with a total batch size of 32, over the course of 500 epochs.

As a post-processing step, we introduce an uncertainty band using dilation to finalize attention to the outputs.

References

[1] A. M. Tahir, M. E. H. Chowdhury, Y. Qiblawey, A. Khandakar, T. Rahman, S. Kiranyaz, U. Khurshid, N. Ibtehaz, S. Mahmud, and M. Ezeddin, "Covid-qu-ex dataset," 2022.

[2] H. Tang, N. Sun, Y. Li, and H. Xia, "Deep learning segmentation model for automated detection of the opacity regions in the chest x-rays of the covid-19 positive patients and the application for disease severity," *medRxiv*, pp. 2020–10, 2020.

[3] MIDRC, "Midrc mrale mastermind challenge: Ai to predict covid severity on chest radiographs," 2023.

[4] C. W. C. C. G. S. J. D. k. L. C. L. P. M. K. M. M. M. P. P. C. S. H. M. T. X. Anouk Stein, MD, "Rsna pneumonia detection challenge," 2018.

[5] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M. E. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, and M. Gabbouj, "Covid-19 infection map generation and detection from chest x-ray images," *Health information science and systems*, vol. 9, no. 1, p. 15, 2021.

[6] H. Luo, Z. Zhou, C. Royer, A. Sekuboyina, and B. Menze, "Devide: Faceted medical knowledge for improved medical vision-language pre-training," *arXiv preprint arXiv:2404.03618*, 2024.

[7] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Communications*, vol. 14, no. 1, p. 4542, 2023.

[8] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.