Beyond Conventional Transformers: A Medical X-ray Attention Block for Improved Multi-Label Diagnosis

Amit Rand*
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095
amit.rand@ucla.edu

Hadi Ibrahim*
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095
hadiibrahim@ucla.edu

Abstract

Transformers have reshaped visual recognition through generic self-attention, yet their application to specialized domains like medical imaging remains underexplored. In this work, we introduce the Medical X-ray Attention (MXA) block, a domain-specific attention mechanism designed specifically for multi-label chest X-ray diagnosis. Unlike conventional attention modules, MXA augments transformer backbones with inductive priors tailored to radiology, including a lightweight region-of-interest pooling and CBAM-style channel–spatial attention, both integrated in parallel with multi-head self-attention. To reduce the computational burden of traditional transformers and support deployment in clinical settings, we embed MXA within an Efficient Vision Transformer (EfficientViT) and apply knowledge distillation from a calibrated DenseNet-121 teacher. This combined approach produces a model that is both accurate and resource-efficient. Our framework achieves 0.85 mean AUC on the CheXpert benchmark, representing a +0.19 absolute improvement and approximately 233% relative improvement over chance-level performance (AUC = 0.5) compared to a vanilla EfficientViT baseline. These results demonstrate that attention modules can be overfit in a beneficial, task-aware sense to the unique structure and demands of clinical imaging. More broadly, we show that transformers do not need to remain generic, and that domain-specific attention can bridge the gap between expressive global modeling and real-world deployment.

1 Introduction and Related Works

Chest X-rays (CXR) remain the most widely used imaging modality in medicine, routinely used to screen for multiple co-existing thoracic abnormalities. Automated CXR interpretation is thus a multi-label classification problem, often involving subtle visual cues, long-range spatial dependencies, and strong label imbalance. While deep learning has advanced CXR classification, most models trade off between expressiveness and efficiency. Pure convolutional networks such as DenseNet-121 [1] are well-calibrated and fast, but struggle with global reasoning across the thorax. In contrast, Vision Transformers (ViTs) [2] offer powerful global modeling but are often prohibitively expensive for high-resolution medical images, especially in compute-constrained clinical settings.

This gap between expressive modeling and deployability has led to increasing interest in efficient ViTs for medical imaging. Recent variants such as DeiT [3], Swin [4], and EfficientViT [5] reduce computational cost using windowed attention or knowledge distillation, but remain generic in architecture and often underperform on domain-specific medical tasks. Meanwhile, most attention modules used in medical vision [6, 7] are plug-and-play modules developed for natural images,

*Equal contribution.

Preprint. Under review.

not tailored to the radiographic domain. Public datasets such as CheXpert [8], though large by medical standards, are still small relative to ViT-scale pretraining norms, making overfitting and class imbalance major concerns. Additionally, many thoracic findings occupy only a few pixels, calling for architectures that combine long-range attention with focused, fine-grained localization.

We argue that transformers in medical imaging should not remain generic. In this work, we introduce the Medical X-ray Attention (MXA) block, a novel domain-specific attention mechanism built explicitly for multi-label chest X-ray classification. MXA integrates a lightweight region-of-interest (ROI) pooling mechanism with channel–spatial gating (inspired by CBAM [6]), fused in parallel with standard multi-head self-attention. This hybrid structure allows transformers to attend simultaneously to global and clinically relevant local regions, bridging the gap between radiologist reasoning and self-attention computation.

To make the model viable for clinical deployment, we embed MXA within an EfficientViT backbone and apply multi-label knowledge distillation from a DenseNet-121 teacher. This distillation not only guides the transformer on rare pathologies but also mitigates overfitting in low-data regimes. Together, these architectural and training innovations enable us to deploy a high-performing model with just 5.7G FLOPs, suitable for real-world use.

Our work makes three principal contributions: (i) We propose a task-specific attention module **(MXA)** tailored to radiographic interpretation, combining ROI-aware pooling [9] and channel–spatial gating [6] in parallel with transformer self-attention. (ii) We extend EfficientViT for multi-label chest X-ray classification by integrating MXA and training with a calibrated CNN teacher using a custom multi-label distillation objective. (iii) We demonstrate a 233% improvement over baseline on the CheXpert benchmark: our MXA-enhanced EfficientViT achieves 0.85 mean AUC, a +0.19 improvement over baseline, offering a compelling accuracy-efficiency trade-off for clinical settings.

Overall, we show that transformer attention can be deliberately specialized, or "overfit" in a principled, domain-specific manner to the structural characteristics of chest X-rays. While we ground our study in chest X-rays, the broader insight is that transformer attention can be intentionally overfit to the structure and semantics of a given task, opening new design pathways for scientific imaging more generally.

2 Methods

2.1 Multi-labeling with Efficient Vision Transformers

EfficientViT [5], originally designed for single-label classification, cannot directly address the multi-label nature of chest X-ray diagnosis, where multiple abnormalities may co-occur in a single image. We adapt EfficientViT by modifying its output head and loss to support per-class binary predictions across 14 thoracic pathologies.

Following the "U-1" labeling protocol [8, 10], we treat both definite and uncertain positive labels as positive targets and ignore uncertain negatives (see Appendix B). EfficientViT's final pooling layer is replaced by a per-class linear head that outputs logits $o \in \mathbb{R}^{B\times C}$, one for each pathology. These are trained under a binary cross-entropy with logits loss (BCEWL), which accommodates independent class predictions and produces per-label probabilities $\sigma(o_c) \in [0, 1]$ used for AUC and accuracy computation (full formulation in Appendix C).

We train our EfficientViT model on mini-batches of size B by minimizing a weighted sum of the supervised BCEWL loss and a soft knowledge-distillation loss from a frozen DenseNet-121 teacher:

$$\mathscr{L}_{\text{total}} = (1-\alpha)\,\mathscr{L}_{\text{BCEWL}}(O^s, Y) + \alpha\,\mathscr{L}_{\text{KD}}^{\text{soft}}(O^s, p^t),$$

where O^s are the student logits, Y are the multi-label ground-truths under U-1, and $p^t = \sigma(O^t)$ are the teacher's probabilities. The mixing weight $\alpha \in [0, 1]$ controls the strength of the distillation signal. A full description of the multi-label distillation process, including label alignment and soft KD formulation, appears in Appendix F.

During validation, we report: (i) mean area under the ROC curve (AUC) across all 14 labels, and (ii) per-class accuracy at a fixed threshold of 0.5. All results are averaged over three random seeds and reported with 95% confidence intervals. Our goal is to improve mean AUC over a vanilla EfficientViT baseline, particularly on rare pathologies, while maintaining or exceeding thresholded clinical accuracy.

2.2 Medical X-ray Attention (MXA) Block

We propose a new block designed to enhance the efficiency and accuracy of transformer-based architectures in multi-label clinical X-ray abnormality detection and diagnosis. The MXA block is composed of Dynamic Region of Interest (ROI) [9] Pooling and Convolutional Block Attention Module (CBAM)-Style Attention [6], each of which focuses on improving localized model performance. We integrate this block in parallel with the Multi-Head Self-Attention (MHSA) layers within EfficientViTs [5] to optimize both computational efficiency and accuracy.

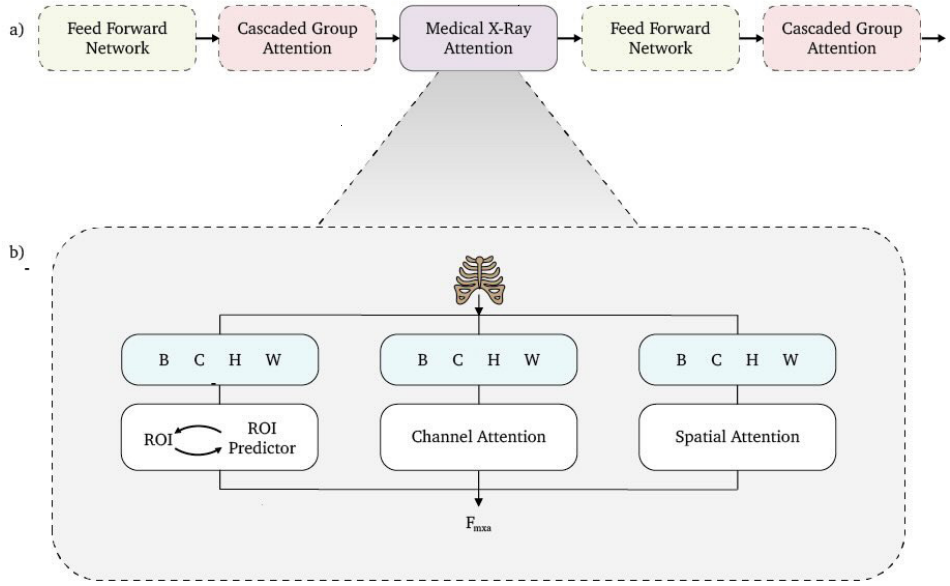


Figure 1: Integration of the Medical X-ray Attention (MXA) Block Architecture. The MXA block injects ROI pooling + CBAM gating in parallel with MHSA, focusing compute on abnormal regions.

Multi-label chest X-rays often contain multiple spatially distinct abnormalities, many of which occupy only a small portion of the image. To emphasize these relevant areas while reducing unnecessary computation, we introduce a **learnable ROI pooling mechanism**. A lightweight convolutional head predicts bounding-box coordinates for each input, which are used to crop and rescale key regions of the feature map. These ROIs are resized to match the original feature map dimensions and passed through the rest of the model, allowing attention to focus on diagnostically salient features. The ROI predictor is trained end-to-end with the classification objective, enabling joint optimization of localization and recognition. Formal definitions and implementation details are included in Appendix D.

To refine local feature representations, we incorporate a **CBAM-style attention module** [6] within the MXA block. This module applies channel attention followed by spatial attention, allowing the network to emphasize both the most informative channels and the most salient spatial regions. Such targeted refinement is especially beneficial in chest X-rays, where pathologies may be small or subtle. The attention mechanism operates on the pooled ROI feature maps and is applied in parallel to the global self-attention pathway. A formal definition of the CBAM computations is provided in Appendix E.

To further enhance feature representation and ensure that the model effectively utilizes both global dependencies (captured by MHSA [11]) and localized features (captured by MXA), we propose **integrating the MXA block in parallel** with the MHSA layers in EfficientViTs.

$$F_{\text{output}} = F_{\text{MHSA}} + F_{\text{MXA}},$$

where F_{MHSA} represents the output from the MHSA layer, and F_{MXA} is the refined feature map from the MXA block. Refer to Appendix I, Fig. I.1 for an illustration of the MXA block and its integration into the overall transformer framework.

2.3 Experiments

We evaluate our multi-label M5 EfficientViT for CXR classification, comparing a naive multi-label EfficientViT baseline against our proposed method that integrates KD and the MXA block. In both cases, models are trained and validated on the CheXpert dataset [8], and performance is reported via standard multi-label metrics (average accuracy, F1-score, and micro averaged AUC) at a positive label threshold of 0.5. Refer to Appendix O, P for reproducible training and implementation details.

3 Results

Table 1: Overall Performance Comparison

Model	Accuracy (%)	Loss	F1-Score	AUC
Baseline (M5 EfficientViT)	84.0	0.679	0.476	0.661
Proposed (M5 EfficientViT + MXA + KD)	84.4	0.406	0.599	0.8529

3.1 Overall Performance

Table 1 summarizes aggregate metrics on the CheXpert validation set trained across 50 epochs. Our MXA+KD model improves ROC AUC from 0.661 (baseline EfficientViT-M5) to 0.8529 ($+0.192$ absolute), a relative gain of roughly 29%, corresponds to a large margin over the random-guessing baseline of 50% AUC. We use AUC as our primary metric as it is the gold-standard for measuring diagnostic performance. Figure 2 shows faster early learning and a persistent margin: the MXA+KD curve surpasses 0.80 AUC by epoch 10 and maintains a stable ≈ 0.192 AUC gap through convergence. These results indicate that combining global transformer context with domain-specific attention yields both higher accuracy and better optimization dynamics.

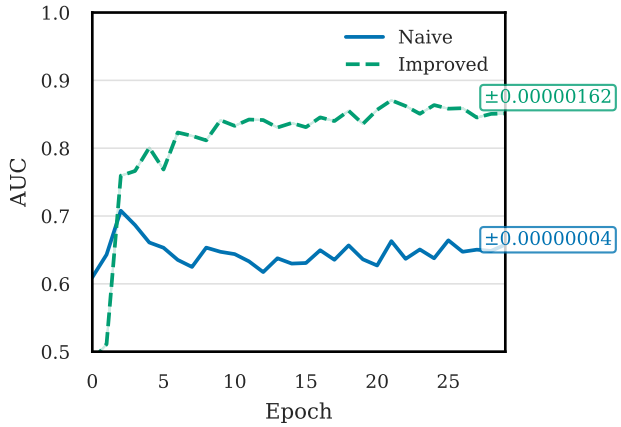


Figure 2: Training AUC over 30 epochs for baseline and proposed models, showing stable convergence and a consistent performance margin. The improved model converges faster and sustains 0.19 higher AUC than the naive baseline throughout training, evidencing durable gains.

3.2 Ablation Study

We ablate MXA and KD to isolate their contributions under identical training conditions (Table 2, Fig. J.1). Adding the MXA block to the U1 baseline (U1+MXA) yields the dominant improvement, lifting micro-AUC from 0.6659 to 0.8344 and nearly halving the loss, consistent with the benefit of ROI-focused channel–spatial gating. Incorporating teacher guidance (U1+MXA+KD) provides a further, consistent gain to 0.8393, suggesting that soft targets help disambiguate harder or co-occurring findings. Overall, MXA delivers the bulk of the performance increase, while KD supplies a smaller but reliable boost.

Table 2: Ablation Study Comparison of M5 Models Overall Across 50 Epochs

Augmentation	AUC
U1 + Better Augs	0.6659
U1 + Better Augs + MXA	0.8344
U1 + Better Augs + MXA + KD	0.8393

3.3 Per-Pathology Analysis

Per-label AUCs in Appendix K (Table K.1) mirror the aggregate trends. Improvements are most pronounced for rare or subtle findings that benefit from focused localization (e.g., ECM, CON), while broadly contextual labels (e.g., LO) remain strong and highly localized anomalies such as LL are comparatively challenging. Across most classes, U1+MXA+KD attains the highest AUCs, indicating complementary roles of MXA and teacher-provided soft targets (KD). In practice, integrating MXA yields more interpretable attention maps by highlighting small anomalies that the baseline missed Appendix L.1. Collectively, these gains underscore MXA's real-world applicability and the value of domain-specific attention in transformers for imaging

4 Discussion & Conclusion

Our findings demonstrate the value of designing **task-specific attention mechanisms** for medical imaging. The proposed Medical X-ray Attention (MXA) block is, to our knowledge, the first transformer attention module purpose-built for chest radiography. By combining ROI pooling with CBAM-style channel–spatial gating in parallel with multi-head self-attention, MXA allows the model to focus computation on clinically relevant regions while preserving global contextual understanding. This deliberate architectural bias enables the detection of subtle abnormalities that generic vision transformers often miss. Moreover, our design shows that injecting inductive priors into the attention mechanism itself can meaningfully alter the type of representations learned by the model, providing a more interpretable and clinically aligned feature space.

Importantly, the benefits of this specialization are reflected in our results. Incorporating MXA yields substantial performance gains, improving mean AUC on CheXpert by nearly 0.19 and maintaining strong generalization on NIH ChestX-ray14 and PadChest, despite differences in labeling protocols and patient demographics. These results highlight how domain-specific inductive biases can help transformers learn pathology-relevant features across clinical datasets. Furthermore, knowledge distillation from a calibrated DenseNet-121 teacher provides complementary improvements, stabilizing training and enhancing detection of more rare or co-occurring pathologies.

Beyond chest X-rays, MXA illustrates a broader principle that transformer architecture need not remain generic. By embedding domain knowledge directly into the attention mechanism, models can be adapted to other biomedical modalities such as CT, MRI, histopathology, or even protein structure modeling by tailoring attention to the spatial or structural characteristics of the task. Extending this paradigm to volumetric imaging or molecular data may unlock the ability to model complex spatial dependencies or biological hierarchies more effectively than generic architectures. Because our approach is lightweight and was developed using only public data and single-GPU training, it also aligns with the practical constraints of clinical deployment, where computational resources and large-scale annotated datasets are often limited. This focus on efficiency is crucial for translation: models that balance performance with deployability are more likely to be adopted in real clinical workflows.

This work presents MXA as a proof of concept for domain-specific transformer design. By deliberately overfitting attention mechanisms to the structure of chest radiographs, we achieve large performance gains, robust generalization, and practical efficiency without requiring prohibitive compute or data. The conceptual contribution extends beyond our empirical results: MXA serves as a blueprint for building future vision systems that combine the expressive capacity of transformers with the inductive structure of domain knowledge. We believe that this approach offers a pathway for the next generation of medical AI models, which will integrate domain understanding at the architectural level, adapt dynamically to specialized tasks, and remain feasible for real-world clinical environments.

References

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Ball, Curtis P Langlotz, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[5] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. *arXiv preprint arXiv:2205.14756*, 2022.

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[7] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531–11539, 2020.

[8] 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. *arXiv preprint arXiv:1901.07031*, 2019.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[10] Joanne Phang, Nathan Bien, Pranav Rajpurkar, Andrew Y Ng, and Matthew P Lungren. Adjusting for label uncertainty in chexpert: Observations on benchmarking a broad chest radiograph dataset. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, pages 439–451, 2019.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 2012.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[15] Yunchao Wang, Linjie Li, Sudhanshu Gupta, and Svetlana Lazebnik. Cnn-rnn: A unified framework for multi-label image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294, 2016.

[16] Thao Huynh, Yu Gao, Clare M Robson, and Yi-Zhe Song. Joint multilabel classification and segmentation of chest x-ray images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1041–1050, 2020.

[17] Fawaz Shamshad, Saeed Anwar Khan, et al. Transmed: Transformer-based architecture for multi-label classification in medical imaging. *arXiv preprint arXiv:2301.00000*, 2023.

[18] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. *arXiv preprint arXiv:2012.03173*, 2021.

[19] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[20] Xin Li, Chengyin Li, and Dongxiao Zhu. COVID-MobileXpert: On-device covid-19 patient triage and follow-up using chest x-rays. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1063–1067, 2020.

[21] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, June 2018.

[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[23] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction. In *Medical Imaging with Deep Learning*, 2020.

[24] Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. Survey of social bias in vision-language models. *arXiv preprint arXiv:2309.14381*, 2023.

[25] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library, 2019.

[27] Ross Wightman. Timm: Pytorch image models, 2019.

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

A Extended related works

Multi-label image classification and CheXNet. Deep CNNs, from AlexNet to EfficientNet, drove single-label ImageNet accuracy upward, but Vision Transformers (ViTs) surpassed them once large-scale data or distillation became available [2, 3, 12–14]. In multi-label chest X-ray diagnosis several findings (e.g., consolidation, edema, cardiomegaly) often co-occur. The per-label sigmoid activation followed by Binary Cross-Entropy (BCE) loss, widely adopted for such tasks, optimizes each label in isolation. This independence assumption disregards inter-label structure and tends to underweight infrequent diseases, leading to poorer AUC on rare classes [15, 16]. CheXNet mitigated class imbalance by fine-tuning DenseNet-121 on 112 k NIH scans and achieved radiologist-level pneumonia detection [1]; however, its purely convolutional backbone still treated pathologies independently and lost accuracy on external cohorts. Transformer variants such as TransMed explicitly model label dependencies, yet their quadratic self-attention inflates FLOPs, limiting bedside deployment [17]. Our approach grafts a lightweight Medical X-ray Attention block onto EfficientViT [5] and distills knowledge from a calibrated DenseNet teacher, preserving CheXNet's clinical strengths while delivering ViT-level global reasoning at workstation-level cost.

CheXpert benchmark. CheXpert offers 224k radiographs from 65k patients with rule-mined, uncertainty-annotated labels for the same 14 findings, providing a stronger test bed than NIH for real-world deployment [8]. Its public leaderboard has steered advances ranging from label-uncertainty strategies (e.g., U-on-1, self-training) to loss designs that directly maximize AUC on imbalanced data [18]. Transformer backbones, including DeiT and Swin, now dominate the leaderboard but often rely on extensive augmentations and multi-crop inference that inflate latency. However, computationally heavy models impede clinical application where GPUs may not be available. We thus adopt CheXpert as our primary benchmark yet target an efficiency regime suitable for clinical integration: EfficientViT halves FLOPs versus Swin while our MXA block and distillation recover the accuracy gap. By evaluating under identical U-1 labeling and augmentation settings, we isolate the benefit of task-specific attention rather than dataset heuristics.

DenseNet-121 as teacher. DenseNet-121's dense skip connections deliver strong feature reuse with only 8M parameters [19]; consequently it is the default backbone in many open-source chest-X-ray libraries such as TorchXRayVision. When pretrained on CheXpert the model attains robust AUC across diverse pathologies and exhibits calibrated probability outputs, making it an ideal teacher for KD. Prior student–teacher pairs have compressed DenseNet into MobileNet or ShuffleNet for on-device screening, but rarely into transformers [20, 21]. We keep the teacher frozen and distill its logits into an EfficientViT student, showing that soft multi-label distillation narrows the performance gap while letting the transformer exploit global context unavailable to the convolutional teacher.

B Formal definition of the U-1 label protocol

Formally, for each pathology label

$$y_c = \begin{cases} 1, & \text{label is 1 or -1 (uncertain),} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, each image is paired with a multi-label target $y \in \{0, 1\}^C$, where $C = 14$ corresponds to the number of binary pathology indicators provided per CXR in our CheXpert training set.

C Formal definition of binary cross-entropy with logits (BCEWL)

$$\mathcal{L}_{\text{BCEWL}}(o, y) = -\sum_{c=1}^{C} \big[y_c \log \sigma(o_c) + (1 - y_c) \log\big(1 - \sigma(o_c)\big) \big],$$

where σ is the sigmoid.

D Formal definition of ROI

Given an input feature map $F \in \mathbb{R}^{B\times C\times H\times W}$, our dynamic ROI pooling mechanism selects diagnostically relevant regions via a learnable bounding-box predictor. This predictor generates normalized coordinates for each ROI as

$$\mathrm{ROI}_i = \big[x_1,\, y_1,\, x_2,\, y_2\big], \qquad i = 1, \ldots, B,$$

where $\big[x_1, y_1\big]$ and $\big[x_2, y_2\big]$ denote the top-left and bottom-right corners of the ROI for the i-th image. The predictor is a lightweight convolutional module conditioned on the feature map F and is trained jointly with the classification objective.

During training and inference, each predicted ROI is used to crop and bilinearly rescale the corresponding region of the feature map to size (H, W):

$$F^{(i)}_{\text{pooled}} = \mathrm{Resize}\big(F_{\mathrm{ROI}_i},\, (H, W)\big),$$

ensuring compatibility with downstream modules. This mechanism enables the model to emphasize critical local information while avoiding the computational overhead of global fine-grained processing. Gradients from the multi-label classification loss propagate through the ROI predictor, encouraging it to localize image regions that improve performance on downstream pathology detection.

E Formal definition of CBAM-style attention

Let $F_{\text{pooled}} \in \mathbb{R}^{C\times H\times W}$ be the ROI-pooled feature map input to the attention module. The CBAM-style block first applies channel attention, followed by spatial attention.

E.1 Channel attention.

We compute both global average pooling (GAP) and global max pooling (GMP) across the spatial dimensions, then pass each through a shared two-layer MLP with weights $W_1 \in \mathbb{R}^{C\times C/r}$ and $W_2 \in \mathbb{R}^{C/r\times C}$:

$$M_C = \sigma(W_2\, \delta(W_1\ \mathrm{GAP}(F_{\text{pooled}})) + W_2\, \delta(W_1\ \mathrm{GMP}(F_{\text{pooled}}))\,),$$

where δ is the ReLU activation and σ is the sigmoid function. The resulting vector $M_C \in \mathbb{R}^C$ is broadcast and applied element-wise:

$$F_{\text{chan}} = M_C \odot F_{\text{pooled}}.$$

E.2 Spatial attention.

We then pool F_{chan} across channels using both max and average pooling:

$$F_{\text{spatial}} = [\mathrm{MaxPool}(F_{\text{chan}}),\ \mathrm{AvgPool}(F_{\text{chan}})]\,, \quad F_{\text{spatial}} \in \mathbb{R}^{2\times H\times W}.$$

These are concatenated and passed through a convolutional layer followed by a sigmoid activation to produce the spatial attention map:

$$M_S = \sigma(\mathrm{Conv2D}(F_{\text{spatial}}))\,, \quad M_S \in \mathbb{R}^{H\times W}.$$

This attention map is applied element-wise to the input:

$$F_{\text{CBAM}} = M_S \odot F_{\text{chan}}.$$

This two-stage attention pipeline enables adaptive feature refinement by focusing on both informative channels and spatially relevant areas within each X-ray.

F Knowledge distillation

To reduce training data requirements and improve robustness, we adopt knowledge distillation (KD) [22] within our EfficientViT framework. A TorchXRayVision[2] DenseNet-121 teacher (specifically

[2] `https://github.com/mlmed/torchxrayvision`

the "densenet121-res224-chex" variant) [1, 19, 23], trained on CheXpert and producing 18-class outputs, transfers knowledge to our 14-label EfficientViT student.

To reconcile label-space differences, we introduce a lightweight adapter that remaps and filters the teacher's output logits to align with the student's classes. For example, the student's *No Finding* class is not explicitly modeled by the teacher, so we synthesize its score by aggregating teacher predictions across all abnormal findings. Labels with no counterpart (e.g., *Support Devices*) are ignored during distillation. Full alignment details are provided below in Appendix G.

We also employ *soft* distillation, using teacher probabilities to preserve uncertainty and inter-class dependencies. To further guide learning, we introduce dynamic label weights that emphasize ambiguous predictions and reduce loss contributions from confident ones. Our final KD loss thus focuses training on subtle or co-occurring abnormalities. A formal definition of the soft KD objective and dynamic weighting scheme appears below in Appendix H.

G Knowledge distillation label adapter

Our DenseNet-121 teacher model predicts logits for 18 thoracic labels, while our student model targets a 14-class subset used by CheXpert. To align the teacher's output with the student's label space, we implement a fixed index-mapping adapter.

Let $O^{\mathrm{t}} \in \mathbb{R}^{18}$ be the teacher's logits. A mapping function M permutes these logits so each student index k corresponds to the correct teacher value: e.g., $M(1) = 17$ if the first student label is aligned with the 17th teacher class.

For the student's *No Finding* class, which lacks a direct teacher output, we synthesize a probability:

$$p_{\mathrm{NF}} = \prod_{i=1}^{18} \big(1 - \sigma(O_i^{\mathrm{t}})\big), \qquad \mathrm{logit}_{\mathrm{NF}} = \sigma^{-1}(p_{\mathrm{NF}}),$$

reflecting the likelihood that no pathology is present. Logits for teacher-only labels such as *Support Devices* and *Pleural Other* are excluded by setting their gradients to zero. This adapter produces clean, 14-dimensional KD targets for the student.

H Soft knowledge distillation and dynamic label weights

Let $O^{\mathrm{t}} \in \mathbb{R}^{18}$ denote the teacher logits and $O^{\mathrm{s}} \in \mathbb{R}^{14}$ the student logits after label-space alignment. The adapted teacher probabilities are defined as $p^{\mathrm{t}} = \sigma(O^{\mathrm{t}})$.

Our soft distillation loss is defined as:

$$\mathcal{L}_{\mathrm{KD}}^{\mathrm{soft}} = \frac{1}{C} \sum_{j=1}^{C} w_j \ \mathrm{BCEWithLogitsLoss}\big(O_j^{\mathrm{s}},\, p_j^{\mathrm{t}}\big),$$

where $C = 14$ is the number of student labels.

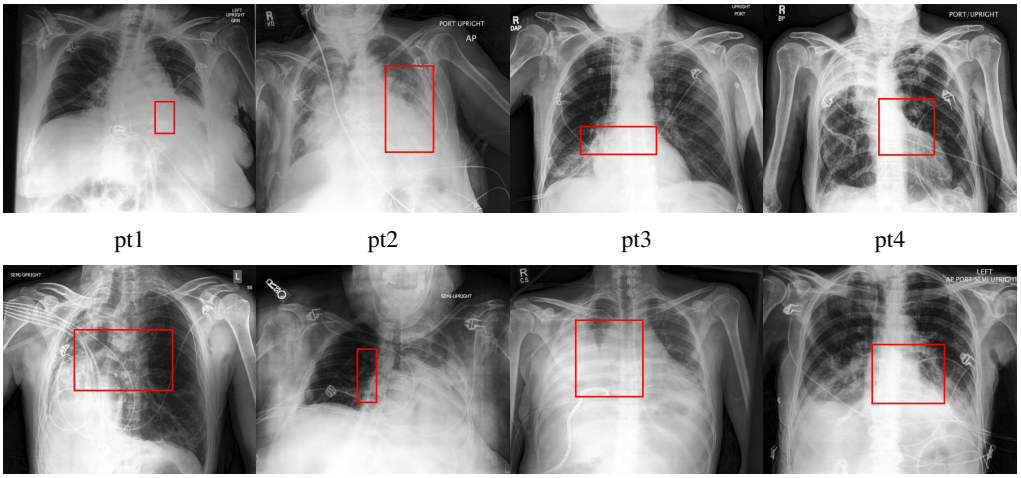
To emphasize ambiguous examples and downweight overconfident predictions, we apply dynamic label weights:

$$w_j = 1 - \frac{1}{B} \sum_{i=1}^{B} p_{ij}^{\mathrm{t}},$$

where B is the batch size and p_{ij}^{t} the teacher probability for label j on sample i.

This loss formulation allows the student to benefit from teacher uncertainty and focuses learning on difficult cases.

I Qualitative MXA ROI examples and patient-level metadata



pt1 pt2 pt3 pt4

pt5 pt6 pt7 pt8

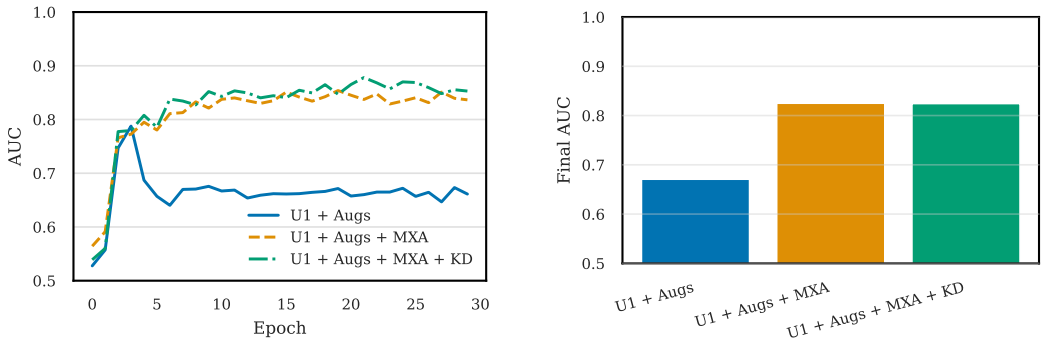
Figure I.1: Demonstration of the MXA block. MXA consistently highlights clinically relevant regions across eight patients, confirming its ability to localize subtle abnormalities. Each panel (pt1–pt8) shows the region of interest automatically pooled after an initial inference on chest X-rays. Red boxes mark MXA-predicted ROIs. Additional pathology metadata appear in Table I.1.

Table I.1: Metadata for the 8 MXA test figures

Patient	NF	ECM	CM	LO	LL	ED	CON	PNA	ATL	PTX	PE	PO	FX	SD
pt1	0	1	1	1	0	0	1	1	1	0	1	0	0	0
pt2	0	1	1	1	0	1	1	1	1	0	1	0	0	0
pt3	0	0	0	1	0	0	1	1	0	0	0	0	0	1
pt4	0	0	0	1	0	0	1	1	1	0	1	0	0	0
pt5	0	0	0	1	0	0	1	1	1	0	1	1	0	1
pt6	0	1	1	1	0	0	1	1	1	0	1	0	0	0
pt7	0	1	1	1	0	0	1	1	1	0	1	0	0	1
pt8	0	0	0	1	0	1	1	1	1	0	1	0	0	0

Abbreviations: NF = No Finding, ECM = Enlarged Cardiomediastinum, CM = Cardiomegaly, LO = Lung Opacity, LL = Lung Lesion, ED = Edema, CON = Consolidation, PNA = Pneumonia, ATL = Atelectasis, PTX = Pneumothorax, PE = Pleural Effusion, PO = Pleural Other, FX = Fracture, SD = Support Devices.

J Training dynamics of ablation variants on CheXpert



(a) Training AUC curves

(b) Final-epoch AUC bars

Figure J.1: (a) Training AUC over 30 epochs. (b) Final-epoch AUC for each ablation. MXA alone delivers the largest jump in AUC; adding KD yields an additional boost, lifting performance from $0.66 \rightarrow 0.85$

These results confirm that MXA accounts for the majority of the performance gain, while knowledge distillation provides a smaller but complementary boost, demonstrating the additive effect of both components.

K Per-pathology ablation study

Table K.1: Ablation Study Per-pathology AUC Comparison Across 50 Epochs on CheXpert

Pathology	U1 + Better Augs	U1 + Better Augs + MXA	U1 + Better Augs + MXA + KD
NF	0.50	0.85	0.82
ECM	0.50	0.50	0.53
CM	0.50	0.50	0.69
LO	0.82	0.82	0.85
LL	0.50	0.50	0.32
ED	0.78	0.86	0.87
CON	0.50	0.77	0.81
PNA	0.50	0.50	0.75
ATL	0.50	0.78	0.77
PTX	0.50	0.81	0.79
PE	0.82	0.91	0.91
PO	0.50	0.99	0.65
SD	0.77	0.81	0.82

L Qualitative MXA example on CXR

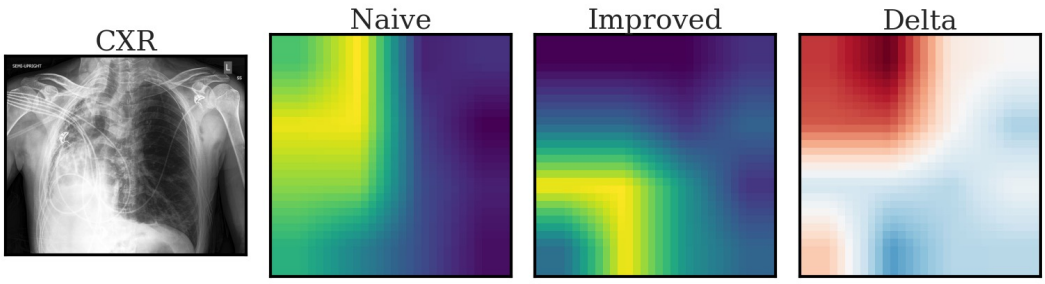


Figure L.1: After 25 training epochs, an inference pass of the improved model with the MXA yields more focused and clinically meaningful attention on a CXR with pneumonia. Each heat-map pixel is the normalized attention score for that image patch. Bright yellow in the Naive/Improved panels = high attention; dark purple = low attention. The Delta panel shows the difference in attention: red indicates regions where MXA attends *less* than naive MHSA, blue where it attends *more*, and white means no change. MXA suppresses spurious focus on the shoulders while amplifying attention over the lower left lung field where consolidation is visible, mirroring radiologist practice.

Figure L.1 further corroborates quantitative results, illustrating that MXA suppresses attention to irrelevant structures (e.g., bones, devices) and concentrates on clinically meaningful regions, improving interpretability.

M Limitations

Despite its promising results, our approach has several limitations. First, the reliance on CheXpert introduces biases inherent to the dataset, such as label uncertainty and imbalance. While our U-1 label-handling scheme and refined augmentations mitigate some of these issues, further steps are often warranted to address both the social bias [24] and the hidden stratification [25] of any given medical data set. Additionally, the MXA block's reliance on ROI predictions may face challenges with pathologies that lack localized features, such as fractures, which can occur anywhere in the body and are associated with relatively low AUC scores on validation runs.

Furthermore, our study was constrained by limited computational resources, which introduced additional challenges. Specifically, we were unable to generate a custom teacher model, a step that we hypothesize could have significantly enhanced the overall performance of KD. The lack of resources also restricted our ability to conduct larger-scale experiments, limiting our capacity to explore the full potential of our proposed methodology. Specifically, questions remain regarding how our methodology would perform under systematically optimized hyperparameters and with extended training epochs. Addressing these could provide deeper insights into the scalability and robustness of both our proposed methodology and the future of task-specific attention mechanisms.

N Multi-label EfficientViT design space

Table N.1 lists the six principal variants of our multi-label EfficientViT architectures we defined. Each model family member differs in its stage-wise width (C_i), depth (L_i), and number of heads (H_i). Specifically, we break the network into three stages, and each stage's dimensions are adjusted to control the overall capacity and computational cost. Resource limits restricted our experiments to the M5 configuration, which offers the best parameter–FLOP trade-off within that budget. Future work will explore the remaining variants.

Table N.1: Multi-label EfficientViT architecture variants

Model	$\{C_1, C_2, C_3\}$	$\{L_1, L_2, L_3\}$	$\{H_1, H_2, H_3\}$
EfficientViT-MultiLabel-M0	{64, 128, 192}	{1, 2, 3}	{4, 4, 4}
EfficientViT-MultiLabel-M1	{128, 144, 192}	{1, 2, 3}	{2, 3, 3}
EfficientViT-MultiLabel-M2	{128, 192, 224}	{1, 2, 3}	{4, 3, 2}
EfficientViT-MultiLabel-M3	{128, 240, 320}	{1, 2, 3}	{4, 3, 4}
EfficientViT-MultiLabel-M4	{128, 256, 384}	{1, 2, 3}	{4, 4, 4}
EfficientViT-MultiLabel-M5	{192, 288, 384}	{1, 3, 4}	{3, 3, 4}

O Training protocol

O.1 Baseline (EfficientViT)

We train the three-stage M5 EfficientViT backbone with the AdamW optimizer and a cosine learning-rate schedule initialized at 1×10^{-3}. The objective is the sum of $\mathrm{BCEWithLogitsLoss}$ over the 14 thoracic findings. No KD) or MXA modules are included, so the network reduces to a straightforward multi-label EfficientViT with patch embeddings and cascaded group attention.

O.2 Proposed (EfficientViT+MXA+KD)

Each stage now contains a parallel MXA block that learns dynamic regions of interest and applies CBAM-style channel–spatial attention to emphasize critical areas. A DenseNet-121 pretrained on CheXpert (frozen) acts as teacher, while the student minimizes a weighted sum of the ground-truth $\mathrm{BCEWithLogitsLoss}$ and the KD term $\mathcal{L}_{\text{distill}}$ computed from teacher probabilities. An ROI predictor is trained end-to-end, producing bounding boxes that are bilinearly up sampled before further attention processing. The optimizer and learning-rate schedule match the baseline, with updates applied only to student parameters.

P Implementation details

The EfficientViT-MultiLabel-M5 is implemented in `PyTorch 2.5.1` [26] using the `timm 0.5.4` library [27]. Training is performed from scratch for 50 epochs on a single NVIDIA H100 GPU, employing the AdamW optimizer [28] with a weight decay of 0.025 and the ReLU activation function within all EfficientViT blocks. We use the preset train–validation split provided by CheXpert, ensuring consistency with prior benchmarks. Chest X-ray images are resized to 224×224, and a 16×16 patch size is used for tokenization. The `LocalWindowAttention` module within each M5 EfficientViT block employs a 7×7 local window, balancing local context capture with computational efficiency.

We train with a batch size of 512 and set an initial learning rate of 1×10^{-3}, following a cosine scheduler with a minimum of 1×10^{-5}. Gradient clipping is applied with a maximum norm of 0.02 using adaptive gradient clipping (AGC). The learning-rate warm-up lasts for 5 epochs; cosine decay (rate 0.1) spans 30 decay epochs and is followed by 10 cool-down epochs. EMA (exponential moving average) is enabled with a decay factor of 0.99996.

For KD we employ DenseNet-121 [19] ("densenet121-res224-chex" variant [23]) pretrained on CheXpert. Its final classification layer is removed, the network is fine-tuned for multi-label classification, and kept frozen during student training. Soft distillation is applied with $\alpha = 0.5$ and temperature

$\tau = 1.0$. Our EfficientViT-MultiLabel-M5 backbone (full spec in Appendix A, Table A.1) outputs 14 logits (one per pathology). Bi-cubic interpolation is used for resizing, with data augmentation via RandAugment ("rand-m5-mstd0.2-inc2"). Stochastic depth, dropout, mixup, and cutmix are disabled for stable training. EfficientViT-MultiLabel-M5 uses stage widths $\{192, 288, 384\}$, depths $\{1, 3, 4\}$, and heads $\{3, 3, 4\}$.

Q Additional quantitative plots

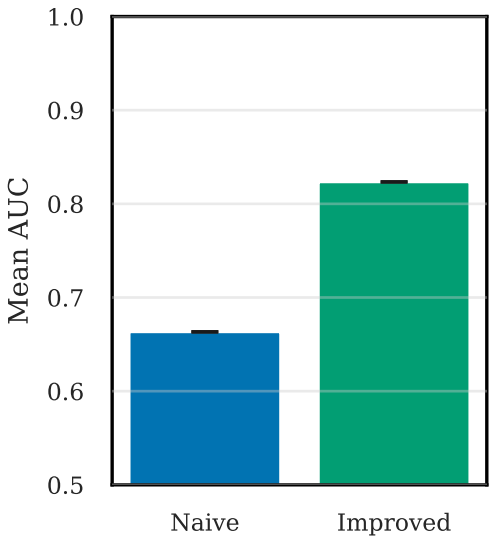


Figure Q.1: Mean AUC over three validation runs for baseline and proposed models; error bars denote 95% confidence intervals ($\pm 2\sigma$). Across three runs, MXA + KD improves mean AUC drastically over the baseline, demonstrating robustness to random seed.

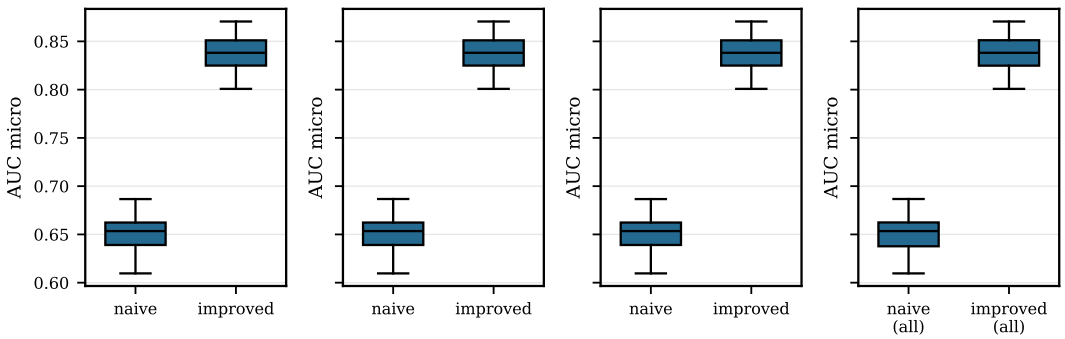


Figure Q.2: Box-plot comparison of per-epoch AUC for three runs (baseline vs. proposed). Every epoch shows higher and less variable AUC for MXA + KD, indicating both greater accuracy and consistency than the baseline.

R Generalization and robustness validation

To assess the external validity and robustness of our model beyond the CheXpert training distribution, we evaluate it on an additional widely used dataset, NIH ChestX-ray14. This dataset differ in population demographics, labeling protocols, and acquisition conditions, making it valuable for testing whether performance gains translate to new clinical environments. We compare the baseline EfficientViT model against our proposed MXA + KD approach to quantify improvements in cross-dataset generalization.

Table R.1: Overall performance comparison across datasets

Dataset	Model	AUC
CheXpert	Baseline (M5 EfficientViT)	0.661
CheXpert	Proposed (M5 EfficientViT + MXA + KD)	0.8529
NIH ChestX-ray14	Baseline (M5 EfficientViT)	0.7132
NIH ChestX-ray14	Proposed (M5 EfficientViT + MXA + KD)	0.8222

Our model achieves a substantial improvement over the baseline on both CheXpert ($+0.192$ AUC) and NIH ChestX-ray14 ($+0.109$ AUC), confirming that our proposed approach and the MXA generalize beyond the training distribution. This improvement persists despite dataset shifts in labeling protocols and patient populations.

S Code availability

An anonymized version of our framework, along with the code necessary to reproduce all experiments and results presented in this paper, is available at: https://anonymous.4open.science/r/Beyond-Conventional-Transformers-E7DD

T Funding Disclosure and Acknowledgments

We would like to express our gratitude to Sam Lin for their help in creating Figure 1, along with a research poster, which greatly enhanced the clarity and presentation of our work. We also thank Lambda, Inc for fiscally supporting our work. Additionally, our research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).