Artificial Intelligence In Medicine 160 (2025) 103054

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

ItpCtrl-AI: End-to-end interpretable and controllable artificial intelligence by modeling radiologists' intentions☆

Trong-Thang Pham [a,*], Jacob Brecheisen [a], Carol C. Wu [b], Hien Nguyen [c], Zhigang Deng [d], Donald Adjeroh [e], Gianfranco Doretto [e], Arabinda Choudhary [f], Ngan Le [a]

[a] AICV Lab, Department of EECS, University of Arkansas, AR 72701, USA
[b] MD Anderson Cancer Center, Houston, TX 77079, USA
[c] Department of ECE, University of Houston, TX 77204, USA
[d] Department of CS, University of Houston, TX 77204, USA
[e] Department of CSEE, West Virginia University, WV 26506, USA
[f] University of Arkansas for Medical Sciences, Little Rock, AR 72705, USA

ARTICLE INFO

Keywords:
Interpretable deep learning
Computer-aided diagnosis
Vision-language model
Radiology
Radiologist's intention
Gaze intention

ABSTRACT

Using Deep Learning in computer-aided diagnosis systems has been of great interest due to its impressive performance in the general domain and medical domain. However, a notable challenge is the lack of explainability of many advanced models, which poses risks in critical applications such as diagnosing findings in CXR. To address this problem, we propose ItpCtrl-AI, a novel *end-to-end interpretable and controllable framework that mirrors the decision-making process of the radiologist*. By emulating the eye gaze patterns of radiologists, our framework initially determines the focal areas and assesses the significance of each pixel within those regions. As a result, the model generates an attention heatmap representing radiologists' attention, which is then used to extract attended visual information to diagnose the findings. By allowing the directional input, our framework is controllable by the user. Furthermore, by displaying the eye gaze heatmap which guides the diagnostic conclusion, the underlying rationale behind the model's decision is revealed, thereby making it interpretable.

In addition to developing an interpretable and controllable framework, our work includes the creation of a dataset, named Diagnosed-Gaze++, which aligns medical findings with eye gaze data. Our extensive experimentation validates the effectiveness of our approach in generating accurate attention heatmaps and diagnoses. The experimental results show that our model not only accurately identifies medical findings but also precisely produces the eye gaze attention of radiologists. The dataset, models, and source code will be made publicly available upon acceptance.

1. Introduction

Deep Learning (DL) has shown remarkable success across a range of fields including Computer Vision [1–3], Natural Language Processing (NLP) [4,5], Autonomous Driving [6,7], and Medical Imaging Analysis [8,9], However, incorporating these advancements into clinical applications poses considerable challenges largely because of their inherently opaque, "black-box" nature. Establishing trustworthiness in clinical applications is crucial, requiring a DL framework that can accurately replicate the decision-making process of actual radiologists. Achieving this level of mimicry poses a significant challenge for current DL models [10].

According to [20–22], the quality of the cue is of the utmost importance in aiding decision-making and thus help to establish trustworthiness from physicians in clinical applications. To gain insights from top-performing black-box diagnostic models, various attempts have been made to interpret these models, notably through Class Activation Mapping (CAM) [12,23,24]. But the attention heatmaps generated from these tools are not entirely reliable because they lack any constraints based on physician-verified ground truth, aside from

☆ Source code and data are available at https://github.com/UARK-AICV/ItpCtrl-AI.

* Corresponding author.
E-mail addresses: tp030@uark.edu (T.-T. Pham), jmbreche@uark.edu (J. Brecheisen), ccwu1@mdanderson.org (C.C. Wu), hvnguy35@central.uh.edu (H. Nguyen), zdeng4@entral.uh.edu (Z. Deng), donald.adjeroh@mail.wvu.edu (D. Adjeroh), gianfranco.doretto@mail.wvu.edu (G. Doretto), achoudhary@uams.edu (A. Choudhary), thile@uark.edu (N. Le).

https://doi.org/10.1016/j.artmed.2024.103054
Received 12 May 2024; Received in revised form 13 October 2024; Accepted 5 December 2024
Available online 12 December 2024

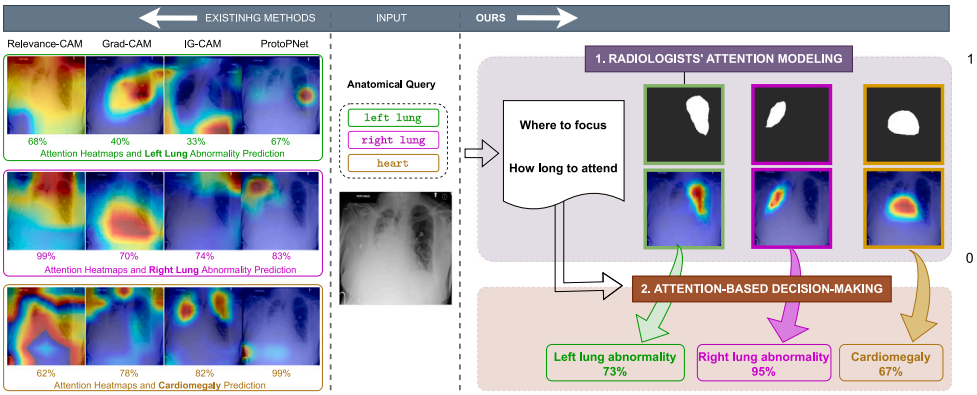

Fig. 1. Comparison between our ItpCtrl-AI and existing DL methods. For the same input and diagnostic query (center), the attention maps and diagnostic decisions from existing DL methods Relevance-CAM [11], Grad-CAM [12], IG-CAM [13], ProtoPNet [14] are displayed on the left whereas the attention maps and diagnostic decisions produced by ItpCtrl-AI, are shown on the right. Although existing DL methods achieve high accuracy in diagnostic decisions, their visual cues do not align with radiologists' attention. On the other hand, ItpCtrl-AI accepts CXR images and queries as input, generating attention masks reflective of radiologist viewpoints ("Where does a radiologist look?") and attention heatmaps indicating the duration of their gaze ("How long do they attend?"). These attention maps are then utilized to make predictions about the presence of abnormalities in each anatomy ("How likely is an abnormality to exist in this area?").

Table 1
Model capacity comparison between ItpCtrl-AI and related Deep Learning approaches. Many approaches provide the ability to tell where visual cues are (localization) with attention weight on pixels (Intensity), and predict the finding (Diagnosis). While many existing methods, such as heatmap-based approaches, provide a degree of interpretability, few closely mimic how a radiologist makes decisions based on what they see. ItpCtrl-AI enhances interpretability by leveraging eye-tracking data from radiologists, allowing it to more accurately emulate the visual attention patterns and decision-making processes of clinical experts. And none offer the user the ability to query which anatomy we want to look at specifically (Controllability). To the best of our knowledge, our method is the first to have all of these attributes.

Methods	Localization	Diagnosis	Intensity	Interpretability	Controllability
CheXNet [15]	✗	✓	✗	✗	✗
Sonsbeek et al. [16]	✗	✓	✗	✗	✗
Grad-CAM [12]	✓	✗	✓	✗	✗
Grad-CAM++ [17]	✓	✗	✓	✗	✗
Relevance-CAM [11]	✓	✗	✓	✗	✗
Integrated Grad-CAM [13]	✓	✗	✓	✗	✗
Rozenberg et al. [18]	✓	✓	✗	✗	✗
Karargyris et al. [19]	✓	✓	✓	✗	✗
ItpCtrl-AI (Ours)	✓	✓	✓	✓	✓

the final disease label. This limitation can lead to the utilization of inaccurate cues, such as interpreting the diaphragm as an indirect indicator for Cardiomegaly [19]. We demonstrate this problem in Fig. 1. Even if we use multiple explainable tools (three CAM methods and one prototype-based method in Fig. 1), they fail to explain why their predictions are correct. This situation necessitates exploring methods to guide the model's focus following radiologists' intention.

We observe that the radiologists heavily rely on their visual skills, carefully examining images to confirm the presence of abnormalities only after gathering sufficient visual information [25]. Therefore, exploring the connection between attention and decision-making can provide valuable insights for reverse-engineering and help good decision-making for reverse-engineering [26]. With a framework based on the attention of skilled individuals, the new decision-makers can learn to process information similarly to skilled individuals, by using extracted visual patterns, and hence they can improve their attention regulation and performance [26]. Despite the importance, extracting meaningful insights from radiologists' attention when diagnosing from eye gaze data remains an open challenge.

Recently, existing works have addressed the issue of localization by making predictions in the form of bounding boxes [18,27]. While these methods predict both the disease and its location using bounding boxes, they are limited in their ability to specify where to focus within the identified anatomy. For example, if the anomaly is simply a thin diagonal line from left to right, and the bounding box covers from its top left point to its bottom right, most of the pixels in that bounding box are irrelevant.

To enhance localization precision, the research community has been increasingly concentrating on segmentation techniques. However, the effective training of segmentation models for gaze attention prediction remains underexplored, as most existing models are designed for anatomical segmentation [28–30]. Aside from UNet and its derivatives [19,31], many segmentation techniques are primarily developed for chest CT images [32,33]. In an attempt to use the gaze information, [19] introduces an eye gaze dataset and modifies UNet [31] to generate both attention heatmaps and predict abnormal findings by using two separate heads on a single bottleneck latent feature. Due to the bottleneck in the design, this model does not address the problem of incorrectly using information for classifications.

Naturally, the simplest form of enforcing a model to use an exact area for classification is to mask out unnecessary pixels. Assume we have enough data, we would want an expert segmentation and classification model for each anatomy. So if we need to perform predictions on three anatomies, we need three models. But in medical applications, localization data is usually rare and limited [34]. This scenario raises a few problems if we train separate models for multiple anatomies. First, the information will be isolated. The model on the left side can only see its left side data. There are many common characteristics that both left and right side of the lungs share, for example, pneunomia can happen on both sides, and its cue is usually the same that the air sacs contain fluid. If there is a finding that is not in the left side data, but we have it on the right side data, this could make the model fail to recognize it. So if we have a way to make the model see both sides of the data, the model has the potential to generalize better in practical use [35]. From these observations, we propose a solution for two problems: (i)

we need a single model to better utilize the limited dataset than existing methods, and (ii) we need a model that can mask out unnecessary information before diagnosing.

To holistically tackle the aforementioned challenges, we propose a novel **unified end-to-end controllable and interpretable pipeline** for simultaneously generating radiologist-based anatomic attention heatmaps and predicting abnormal findings. We design our model with the objective to mimic how a radiologist makes a decision: observe visually and then make a diagnosis based on what is seen. In other words, our system solves two challenges: Radiologists' Attention Modeling and Attention-based Decision-making. As illustrated in Fig. 1, our method utilizes short textual prompts that specify an anatomical area. These prompts guide the model in generating corresponding heatmaps and masks, offering us the flexibility to choose the region of interest. By using prompts, we can have one model to do multiple tasks, which addresses the aforementioned problem (i), and gives us the *controllability* characteristic. To force the model not to randomly generate attention and learn how the radiologist looks into the CXR, we constrain it with two direct objectives (Section 2) from the ground truth eye gaze data. By understanding where and with what intensity the radiologist focuses, we can filter out irrelevant data before the classification step. This ensures that our model does not rely on incorrect information, as follows how the radiologist would do, which addresses the aforementioned problem (ii). As a consequence, we can understand the decision of the model because each diagnosis is paired with an attention heatmap, which gives us the *interpretaility* characteristic. As a result, our system stands apart from existing black-box models by offering greater interpretability. Users can glean meaningful insights from each module of our system independently [36]. Furthermore, this design allows our model to utilize all information from the datasets. Table 1 illustrates the overall capacity of our model compared to existing methods.

To the best of our knowledge, there are no public datasets that associate radiologists' anatomic attention heatmap with each finding. To address this, we propose the DiagnosedGaze++ dataset and its semi-automatic curated procedure. To obtain the radiologist's attention intensity, we use the REFLACX dataset [34], which contains a plethora of eye gaze information captured by high-sensitivity hardware of radiologists analyzing CXR images. The original data only provides the raw eye gaze with noise. Aligning this gaze data with specific findings is complex due to the dynamic nature of gaze attention. For instance, a radiologist's focus may rapidly switch between the heart and various lung sections before making a diagnosis. This variability makes it difficult to pinpoint the exact gaze points crucial for diagnosis. Given the complexity of gaze attention heatmaps and their alignment with abnormal findings, manual annotation is challenging. To address this, we introduce a semi-automatic method that filters gaze data based on lung anatomy, resulting in a curated eye gaze dataset with gaze attention for each part of the lung, including the corresponding abnormality ground truth.

Our contributions can be summarized as follows:

- We introduce *a novel end-to-end controllable & interpretable approach*, called ItpCtrl-AI, that uses a CXR image in conjunction with an anatomical prompt to determine the location and intensity of the radiologist's focus followed by the prediction of a corresponding finding. To the best of our knowledge, our method is the first in the medical domain to learn from radiologist-based anatomical gaze attention while offering controllability.
- We introduce DiagnosedGaze++, a gaze dataset designed to be interpretable for CXR diagnostic purposes. To create this dataset, we employ a semi-automated method that utilizes transcripts and anatomic gaze masks to extract heatmaps based on radiologists' eye gaze data.
- We performed extensive experiments to validate the effectiveness of ItpCtrl-AI. To ensure reproducibility, we will release the source code, annotated dataset, and trained models upon acceptance.

This journal version significantly extends its conference predecessor [37]. First, we introduce an end-to-end framework instead of the two-stage framework in [37], thereby advancing the method of modeling radiologists' intentions for CXR analysis. Secondly, we double the sample size of our proposed dataset. Then, we enrich the comparative analysis with other segmentation methods, validating the efficacy superiority of our method over existing ones. Finally, we broaden the scope with extensive ablation studies to dissect the contribution of each component in our model. This work significantly enhances the original [37] by offering a more robust, interpretable, and controllable AI system for medical image diagnosis.

2. Related work

Explainable Deep Learning. Understanding a model's decision-making process holds significant importance today, particularly in computer-aided diagnosis (CAD) systems. Recent developments such as Class Activation Mapping (CAM) [11–13,17] have showcased one common approach: training a black box model and subsequently employing CAM-related techniques to visualize critical areas. While black-box models often exhibit high performance, they are recognized for their unreliability, as highlighted in literature [38]. Our work diverges from traditional black-box techniques and their reliance on post-hoc visualization to interpret black-box models. Unlike these methods, which are prone to unreliability despite their high performance, our model is built with interpretability at its core. We aim to closely replicate the diagnostic approach of the radiologist, aiming to make the model's decision-making process transparent from the outset rather than relying on after-the-fact explanations.

Interpretable Deep Learning. Unlike the aforementioned explainable tools, a more desirable approach entails the design of a system wherein decisions are intrinsically linked to explainability, particularly in high-stakes medical contexts [38]. Generally, interpretable models aim to transform inputs into human-interpretable representations such as concepts or prototypes, which are then harnessed for prediction. To imbue the model with self-explanatory capabilities, many researchers have embraced successful prototype-based approaches [14,39–41]. For example, ProtoPNet [14] introduces a prototypical part network that identifies prototypical parts within input images, leveraging this insight for the final prediction. PIP-net [41] learns prototypes that align closely with human visual perception, serving as scoring sheets during classification. However, these approaches rely on the automatic learning of prototypes to explain predictions, which can result in prototypes that are not intuitively understandable by humans. Notably, TCAV [42] is trained to identify important concepts from a user-defined set of concepts when predicting a class, such as recognizing important "striped" to classify the type of "zebra". Yet, its reliance on categorical data for training and its concept importance scoring can hinder the exploration of less overt concepts, such as gaze attention. Our approach extends beyond these methodologies by utilizing radiologist-annotated gaze data to guide the model to directly learn gaze attention and produce attentions that reflect the radiologist's perspective. This direct inclusion of gaze data sets our work apart by providing a realistic layer of interpretation grounded in actual radiological practice.

Disease Condition Localization. Some existing methods [18,27,43,44] predict a bounding box to localize diseases, with the ground truth being a bounding box and a disease label. However, the resulting bounding boxes usually include significant portions of irrelevant space [31]. Other works [45,46] train the model primarily on an image-level label and extract saliency maps or use Class Activation Mapping (CAM) to obtain the location of the disease. However, these works lack constraints on where the model should look, which results in random, uninterpretable attention. The work of [19] predicts an attention, but the ground truth is a full gaze map, and its GradCAM visualization indicates that the model is unreliable as it incorporates

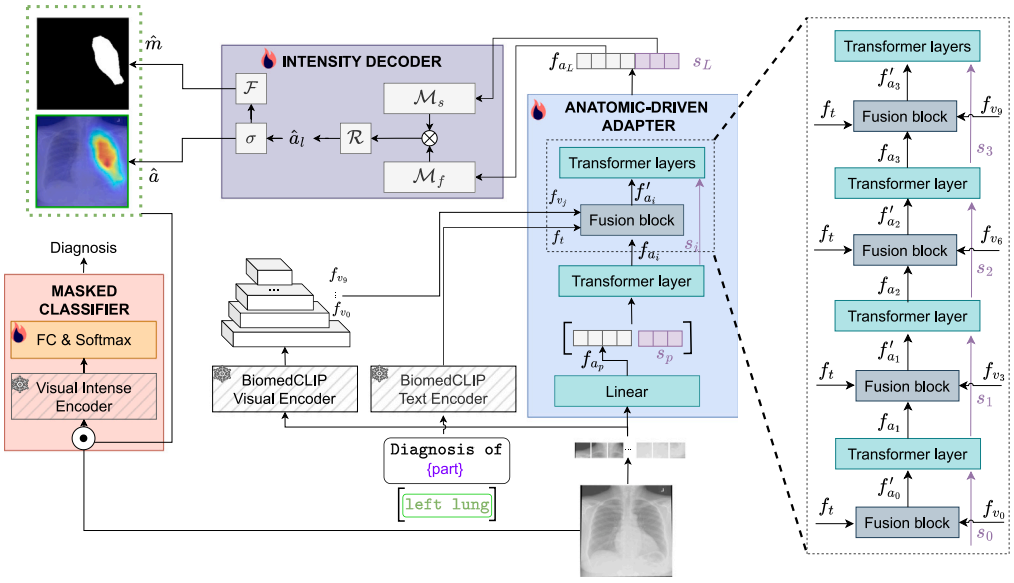

Fig. 2. The detailed pipeline of our proposed controllable & interpretable framework ItpCtrl-AI to decode radiologists' intense focus for accurate CXR diagnoses. ⊙ is the Hadamard product.

unrelated information in classification and attention prediction. By contrast, we utilize a unique combination of eye-tracking information, the reading of the radiologist, and anatomic segmentation to generate precise anatomic radiologist-based attention.

CXR Disease Classification. Disease classification using CXR images has gained much attention recently. The earliest of these efforts, ChexNet [15], is a DenseNet [47] that uses the entire CXR image as its direct input. Since then, many efforts that use deep learning have risen from the related areas of supervised learning [48,49], semi-supervised learning [50,51], and self-supervised learning [52,53]. Besides using the whole CXR to predict disease, numerous studies [16,27,54–58] suggest that location information of the disease can help in classification tasks. However, the above methods generally include irrelevant areas, such as some pixels within bounding boxes, or they learn automatically without any constraints that align with real-world explanations, such as with prototypes or CAM. To the best of our knowledge, no existing methods use anatomic radiologist-based attention in aiding and masking out irrelevant pixels in medical images for classification.

3. Problem formulation

Given a CXR image x and an anatomical query q, e.g., "`Diagnosis of {}`" as a prefix with "`left lung`", "`right lung`", or "`the heart`", our goal is to produce a radiologist-based attention heatmap a, gaze mask m, and the corresponding finding y.

A generated attention heatmap and gaze mask must be close, both location-wise and intensity-wise, to the eye gaze attention of the radiologist, constructed in Section 5. This ensures that the generated attention accurately captures the radiologists' attention heatmaps and provides meaningful insights into their diagnostic process. In addition, the predicted finding y should only be based on the similarity of visual information that the radiologist would use in practice.

4. Architecture

Our model tackles the formulated problem in Section 3 by introducing ItpCtrl-AI, consisting of three main modules: *Anatomic-Driven Adapter*, *Intensity Decoder*, and *Masked Classifier*, where *Anatomic-Driven Adapter* and *Intensity Decoder* are used to solve the Radiologists' Attention Modeling challenge, and *Masked Classifier* solves the Attention-based Decision-making challenge, as described in Fig. 1. First, our model takes a CXR image x with the size of $H \times W$ and an anatomical query q as the inputs and feeds them into a visual encoder and text encoder from a medical-specialized CLIP model, called BiomedCLIP [4]. Then, we propose a novel end-to-end lightweight *Anatomic-Driven Feature Extractor* module comprising novel adapter blocks to fuse these encoded features into one feature. Then, we feed the fused features into the *Intensity Decoder* to produce the attention and gaze mask. Finally, we feed all attention, gaze mask, and the original image into the *Masked Classifier* to predict finding y. The architecture is illustrated in Fig. 2.

4.1. Pretrained feature encoders

To utilize a strong existing CLIP model in medical imaging, we leverage a pretrained BiomedCLIP [4], which was trained on 15 million image-caption pairs in PMC-15M [4], to adapt with the small size of our dataset. BiomedCLIP has two encoders: Visual Encoder and Text Encoder. The Visual Encoder is a Vision Transformer (ViT) model [1] that uses patches with 16×16 pixels in size and has 12 transformer layers. The Text Encoder is a BERT model, called PubMedBERT, and has 12 transformer layers. As the pretrained CLIP model is already well-trained, we only need a lightweight adapter to adapt to our task, proposed in Section 4.2

BiomedCLIP Visual Encoder. We are inspired by [59] to use the intermediate features of the Visual Encoder of BiomedCLIP as it provides more useful information than only using the features of the final layer. First, we feed the image x into the BiomedCLIP Visual Encoder and extract the intermediate features $f_{v_j} \in \mathbb{R}^{H/16 \times W/16 \times 768}$ from 4 layers, i.e., 0, 3, 6, and 9, where $j \in \{0, 3, 6, 9\}$ and we count with 0-based index.

BiomedCLIP Text Encoder. Unlike visual encoding, the anatomical prompts are short, i.e. one or two words, and we use them to instruct the direction that the model should focus. Therefore, we only obtain the final embedding $f_t \in \mathbb{R}^{512}$ from the BiomedCLIP Text Encoder module.

4.2. Anatomic-driven adapter

Our Anatomic-Driven Adapter (ADA) is a deep adapter with architecture based on $(L+1)$-layer Vision Transformer (ViT) [1], which leverages a Vision Transformer (ViT) architecture to process chest X-ray (CXR) images by embedding image patches, applying transformer

Algorithm 1 Anatomic-Driven Adapter (ADA) Pipeline. $[\cdot,\cdot]$ is the concatenation operation.

Input: CXR image x
Split x into multiple 16×16 patches
$f_{a_p} \leftarrow \text{Linear}(\text{patches})$, where $f_{a_p} \in \mathbb{R}^{(H/16 \times W/16) \times D}$
Initialize $s_p \in \mathbb{R}^D$ as a scaling vector.
$(f_{a_0}, s_0) \leftarrow \text{TransformerLayer}_0([f_{a_p}, s_p])$
for $i = 0$ to $L - 1$ **do**
  $f'_{a_i} \leftarrow \text{Fusion}(f_{a_i}, f_{v_i}, f_t)$
  $(f_{a_{i+1}}, s_{i+1}) \leftarrow \text{TransformerLayer}_{i+1}([f'_{a_i}, s_i])$
end for
Finally, we obtain the last layer feature f_{a_L} and the scaling vector s_L

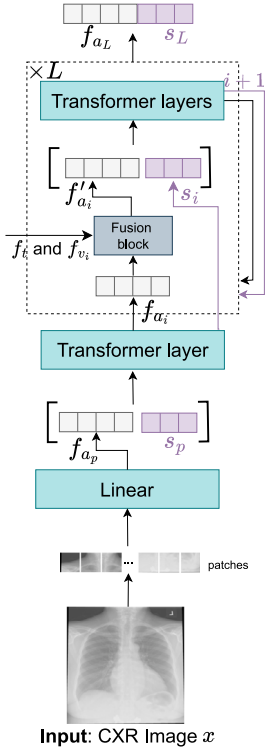

Fig. 3. Illustration of Anatomic-Driven Adapter (ADA) Pipeline.

layers, and fusing features to produce final layer features and scaling vectors. The detailed pipeline is shown in Fig. 3 and Algorithm 1.

Scaling vector. The scaling vector plays an important role in generating gaze attention in Section 4.3. The intuition of including the scaling vector is that each element in the last latent feature does not contribute equally across all anatomic parts, so the learnable scaling vector allows the model to flexibly re-weight the last feature in the most suitable way to produce the final intense attention. By concatenating f'_{a_i} with s_i into an input before feeding into transformer layers, the last scaling vector s_L can gain important information of f_{a_i}, where $i \in [0, L]$, thanks to self-attention operation.

Fusion blocks. The fusion block has three inputs: the BiomedCLIP visual encoding at the jth block $f_{v_j} \in \mathbb{R}^{H/16 \times W/16 \times 768}$, the BiomedCLIP text embedding $f_t \in \mathbb{R}^{512}$, and the adapter latent feature $f'_{a_i} \in \mathbb{R}^{(H/16 * W/16) \times D}$ at the ith block. Note that, this fusion block fuses only the features of the input image, so we do not use a scaling vector as it is used for another purpose.

Mathematically, we obtain the fused adapter feature f'_{a_i} by:

$$f'_t = \text{Linear}_{512 \to D}(f_t), \tag{1}$$

$$f'_{v_j} = \text{Flatten}(\text{Conv2d}_{s=1,k=1,out=D}(f_{v_j})), \tag{2}$$

$$f'_{a_i} = f_{a_i} \oplus f'_{v_j} \oplus f'_t, \tag{3}$$

where operator $\oplus$ is the element-wise adding operator, $\text{Linear}_{512 \to D}(\cdot)$ is the Linear layer with input dimension of 512 and output dimension of D, $\text{Conv2d}_{s=1,k=1,out=D}(\cdot)$ is the convolutional layer with stride (s) of 1, kernel size (k) is 1×1, and the output channel size is D, and $\text{Flatten}(\cdot)$ is the flatten operation to change a 2D matrix of size $H/16 \times W/16$ to a vector of size $H/16 * W/16$. Note that f'_{v_j} and f_{a_i} have the same shape of $(H/16 * W/16) \times D$, but f'_t's dimension is D, so the add operation of f'_t broadcasts across the first dimension of f'_{v_j} and f_{a_i}.

To simplify the pipeline, we only use the add operation for feature fusion because it is a simple and strong established baseline [59]. While other fusion mechanisms may enhance performance, they are beyond the scope of this paper.

4.3. Intensity decoder

Our Intensity Decoder produces two main outputs: the gaze attention $\hat{a} \in [0,1]^{H \times W}$ and gaze mask $\hat{m} \in \mathbb{B}^{H \times W}$, where the set $\mathbb{B} = \{0, 1\}$. This Intensity Decoder module answers "where to focus" (location) and "how to attend" (intensity) by using the features from the Anatomic-Driven Adapter.

Creating gaze attention. The Intensity Decoder receives the output of the last layer of our adapter, i.e. latent feature $f_{a_L} \in \mathbb{R}^{(H/16 * W/16) \times D}$ and the scaling vector $s_L \in \mathbb{R}^D$, to generate the attention heatmap. We first pass f_{a_L} and s_L into two separated multilayer perceptrons (MLPs) $\mathcal{M}_f$ and $\mathcal{M}_s$, respectively, to project both vectors into another latent space for decoding intensity. We then use matrix multiplication between them to produce a small gray-scale attention logit $a'_l \in \mathbb{R}^{(H/16 * W/16)}$. To get the final attention logit, we resize a'_l into $\hat{a}_l \in \mathbb{R}^{H \times W}$. Finally, we obtain the gaze attention by applying a sigmoid function.

Mathematically, we obtain the predicted gaze attention $\hat{a}$ by

$$a'_l = \mathcal{M}_f(f_{a_L}) \otimes (s_L), \tag{4}$$

$$\hat{a}_l = \mathcal{R}_{(H/16 * W/16) \to (H \times W)}(a'_l), \tag{5}$$

$$\hat{a} = \sigma(\hat{a}_l), \tag{6}$$

where $\otimes$ is the matrix multiplication operation, $\mathcal{R}_{(H/16 * W/16) \to (H \times W)}(\cdot)$ is an operator that first reshapes a vector of size $H/16 * W/16$ to $H/16 \times W/16$ and then resizes it to a 2D matrix of size $H \times W$, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function to normalize the value range to 0 and 1 that matches our ground truth heatmap a.

Creating gaze mask. Given the predicted attention mask $\hat{a} \in [0,1]^{H \times W}$, we create the predicted mask $\hat{m}$ by applying a step function $\mathcal{F}(\cdot)$ with a threshold of 0.5:

$$\mathcal{F}(x) = \begin{cases} 1 & \text{if } x \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

In other words, we have

$$\hat{m} = \mathcal{F}(\hat{a}). \tag{8}$$

4.4. Masked classifier

Finally, the Masked Classifier tells us whether the input CXR is abnormal or not. Using the predicted attention $\hat{a} \in [0,1]^{H \times W}$ and mask $\hat{m}$ from the previous step, we use Hadamard product on $\hat{a}$ and $\hat{m}$ with the image x to re-weight the importance of all pixels, called x'. Afterward, we pass x' to our Visual Intense Encoder ($\mathcal{V}$), a Fully Connected (FC) layer followed by Softmax activation to extract and produce the finding probability $y' \in [0, 1]$. In our implementation, the Visual Intense Encoder is the same frozen BiomedCLIP Visual Encoder as in Section 4.2 to avoid unnecessary parameters. Formally, we obtain the probability y' by

$$x' = \hat{a} \odot \hat{m} \odot x, \tag{9}$$

$$y' = \text{softmax}\, FC(\mathcal{V}(x')), \tag{10}$$

where $\odot$ is the Hadardmard product. At inference time, we obtain the predicted finding $\hat{y}$ by applying a threshold of 0.5, i.e. $\hat{y} = \mathcal{F}(y')$.

4.5. Losses

Gaze attention loss $\mathcal{L}_h$. Given the predicted logit $\hat{a}_l \in \mathbb{R}^{H\times W}$ and ground truth gaze attention $a \in [0,1]^{H\times W}$, we compute the gaze attention $\mathcal{L}_h$ loss as L_2:

$$\mathcal{L}_h = \|\hat{a}_l - \sigma^{-1}(a)\|_2, \tag{11}$$

where $\sigma^{-1}(x) = \ln \frac{x}{1-x}$ is the logit function. Note that we compute the loss before applying the sigmoid function to the predicted logit attention to avoid the issue of vanishing gradients.

Gaze mask losses $\mathcal{L}_m$. Given ground truth gaze mask $m \in \mathbb{B}^{H\times W}$, where the set $\mathbb{B} = \{0,1\}$, we compute the gaze mask loss $\mathcal{L}_m$ by combining the standard binary cross entropy loss L_{bce} and dice loss L_{dice} on the mask probability, i.e. attention, $\hat{a}$ and mask ground truth m as in [60].

$$\mathcal{L}_m = L_{bce}(\hat{a}, m) + L_{dice}(\hat{a}, m). \tag{12}$$

Classification loss $\mathcal{L}_c$. We use the standard binary cross entropy loss between the predicted probability finding y' and the ground truth finding y to guide the classifier. Formally, we have

$$\mathcal{L}_c = L_{bce}(y', y). \tag{13}$$

Finally, we train the architecture with the final objective as

$$\mathcal{L} = \mathcal{L}_h + \mathcal{L}_m + \mathcal{L}_c. \tag{14}$$

5. DiagnosedGaze++ dataset

REFLACX [34] provides eye gaze data for more than 2500 CXRs from MIMIC-CXR [61], where each gaze sequence is captured using a device with a sensitivity of 1000 Hz. However, REFLACX does not provide a gaze map for each anatomic part of the lung. With how random each gaze sequence is, we manually annotated the data to construct the disease-level gaze attention heatmap ground truth. The process of creating the ground truth is discussed in Sections 5.2 and 5.3. An overview of our data processing is shown in Fig. 4. The Raw Input panel on the left shows an original chest X-ray and a chest X-ray with overlaid eye-tracking data, indicated by white dots. These dots represent the gaze points of a radiologist examining the X-ray. The annotations on the left side of the image provide timestamps (in second) and corresponding observations made by the radiologist, such as "0–4 s: <say nothing>" and "8–15 s: right basilar opacity may represent consolidation". The second panel, called Filtered gaze on keywords, displays the filtered gaze paths corresponding to specific keywords mentioned by the radiologist, such as "left", "right", and "heart". These paths are color-coded to match the keywords, and the images are filtered to show only the gaze data relevant to those terms. The slider positioned above the images denotes the timestamp value, which signifies the end time of the interval for the gaze points corresponding to each keyword: the "left" gaze timestamp commences at 0 and concludes after 8 s, the "right" gaze timestamp begins at 0 and terminates at 15, and the "heart" gaze timestamp starts at 0 and ends at 17. In the panel below, the original chest X-ray is segmented into three anatomical structures: the left lung, right lung, and heart by using a finetuned SAMed. We apply Hadamard product $\odot$ on the keyword-filtered gaze and the anatomical segmentation masks to produce the final gaze sequence, as shown in the top right panel. The bottom right row of the figure presents two outputs of the eye gaze data: gaze attention and gaze mask.

Table 2
Data distribution corresponding to four distinct settings: C: cardiomegaly, L: Left lung, R: Right lung, M: entire chest and merging all samples from the C, L, and R subsets.

Settings	No. samples (train:val:test)
C	626:117:137
L	1458:313:308
R	1311:297:268
M	3395:727:713

Table 3
Keywords list.

Categories	Keywords
C	''cardiomegaly''; ''enlarged'' and ''cardiac''; ''enlarged'' and ''heart''; ''normal'' and ''heart''; ''cardiomediastinal''; ''mediastinum'' and ''normal''.
L	''left''
R	''right''
M	C ⋃ L ⋃ R

Note that the only category that has more than 300 samples after annotation is Cardiomegaly. Therefore, Cardiomegaly is treated as a separate subset, while all other diseases are categorized into left or right lung subsets. After labeling the data, we split it into four distinct settings:

- C: Only samples with the transcript that specifically mentions cardiomegaly.
- L: Only samples with the transcript that specifically mentions left lung.
- R: Only samples with the transcript that specifically mentions right lung.
- M: Merging all samples from C, L, and R.

For each subset, we split 70% for training, 15% for evaluation, and 15% for testing. We also keep the balance between the positive and negative ratios to be 1:1. The data distribution is shown in Table 2.

5.1. Keywords

The annotation transcripts are created by five radiologists with different word usage for one finding. For example, some may say `enlarged cardiac silhouette`, while others say `cardiomegaly`. So to address the variability, we will take all radiologist's verbal transcripts into consideration. For each sentence, we determine the keywords to decide its category. The selected keywords for each category are as follows (see Table 3).

Some example sentences for each category:

- C: marked cardiomegaly is present; heart size is large; cardiac silhouette is enlarged.
- L: possible small left pleural effusion; left basilar subsegmental atelectasis; no consolidation in the left lung.
- R: opacity within the basilar right hemithorax with curvilinear superior margin may represent a subpulmonic pleural effusion parenchymal opacity; right basilar pneumothorax; right lung is clear.

5.2. Classification

Based on our four distinct settings, we design four yes/no questions for classifying findings.

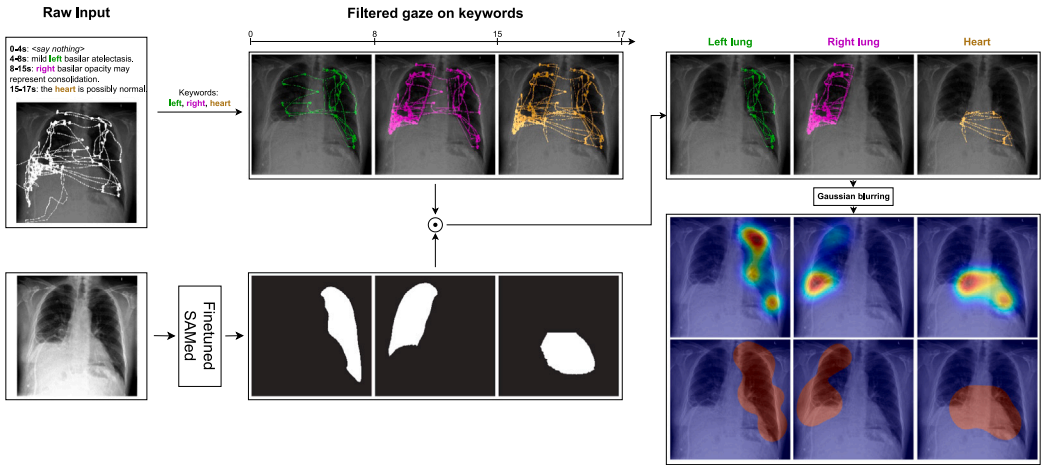

Fig. 4. The pipeline of creating ground truth gaze attention heatmaps and gaze masks of DiagnosedGaze++.

Table 4
Abnormal findings in each setting.

Settings	Findings
C	Cardiomegaly
L	Pleural abnormality, Enlarged hilum, Consolidation, Pulmonary edema, Atelectasis, Lung nodule or mass, Ground glass opacity, Interstitial lung disease, Pneumothorax, Emphysema.
R	Pleural abnormality, Enlarged hilum, Consolidation, Pulmonary edema, Atelectasis, Lung nodule or mass, Ground glass opacity, Interstitial lung disease, Pneumothorax, Emphysema.
M	C ⋃ L ⋃ R ⋃

- C: Is there cardiomegaly?
- L: Is there a finding (excluding Cardiomegaly) in the left lung of the image?
- R: Is there a finding (excluding Cardiomegaly) in the right lung of the image?
- M: Is there a finding in the masked image?

Table 4 shows the abnormal findings in each setting. For each sample of a particular setting, the label is **True** if at least one of the setting's findings is **True**, otherwise **False**.

5.3. Ground truth attention heatmap

To create the ground truth attention heatmap, we perform two steps: make anatomic masks and filter fixations. Fig. 4 demonstrates the overall pipeline for making ground truth gaze attention heatmaps.
Anatomic masks. REFLACX [34] also does not provide anatomic masks, so we have to create these masks as well. Currently, the anatomic masks for three big parts are provided by EGD-CXR [19]: left lung, right lung, and the mediastinum. We finetune SAMed [62] on EGD-CXR, then use the finetuned model to make inferences on REFLACX. Then, we manually correct the gaze masks if there is any problem. For example, we heuristically cut out the top half of each mediastinum mask to make the heart masks, but the automatic script may cut too much, so we check and fix it to correctly cover the heart.
Filtering fixation sequence. For a particular anatomic region, we can acquire the fixations by looking for keywords (see Section 5.1) in the sentences of the provided transcript. For example, `cardiomegaly` or `enlarged` and `cardiac` for setting C. Then, we will pick the rightmost sentence to decide the upper end of the interval containing our desired fixations. Specially, given a sequence of sentences $\{s_1, s_2, \ldots, s_n\}$, if we find s_3, s_4 and s_{10} contain the keyword, we will use s_{10}. As a result, the chosen fixations are in the interval $[0, e]$, where e is the ending time of the sentence s_{10}. Using the predicted mask from before, we exclude any fixation point located beyond its boundaries. Note that the starting time of 0 is required to capture potentially relevant visual information from the moment the radiologist takes their very first glance. Finally, by applying a Gaussian filter with a radius of 150 on the chosen fixations' coordinates [19], we obtain the final ground truth attention heatmap.
Anatomical prompt. We also need the input prompt to guide the model. For our anatomical prompt, we use the prefix "`diagnosis of {}`".

After the prefix, we append our target: "`left lung`" for left lung heatmap prediction, "`right lung`" for right lung heatmap, and "`heart`" for heart heatmap.

Table 5
Dataset comparison between existing CXR datasets and our DiagnosedGaze++.

Dataset	Classification	Gaze data	Anatomy-aware gaze
Chest ImaGenome [63]	✓	✗	✗
REFLACX [34]	✓	✓	✗
EGD [19]	✓	✓	✗
DiagnosedGaze++	✓	✓	✓

5.4. Ground truth gaze mask

Given a ground truth attention $a \in [0, 1]^{H \times W}$, we create the ground truth gaze mask m by applying the step function $\mathcal{T}(\cdot)$ (Eq. (7)) on all pixel values of a to create the ground truth mask m.

6. Experimental details

6.1. Datasets

Table 5 compares the advantages of different datasets in terms of their support for classification tasks, availability of gaze data, and presence of anatomy-aware gaze annotations. The Chest ImaGenome dataset supports classification tasks but lacks gaze data and anatomy-aware gaze annotations. The REFLACX and EGD datasets both support classification tasks and include gaze data, but do not provide anatomy-aware gaze annotations. In contrast, the DiagnosedGaze++ dataset supports classification tasks, includes gaze data, and uniquely offers anatomy-aware gaze annotations, making it the most comprehensive dataset among those listed for these specific features.

We use two datasets for our experiments: DiagnosedGaze++ and Chest ImaGenome dataset [63]. DiagnosedGaze++ provides anatomy-aware gaze, offering a unique evaluation aspect. The Chest ImaGenome dataset, derived from the public MIMIC-CXR [61], is a detailed collection of chest X-ray images annotated with comprehensive anatomical and pathological labels. It comprises over 240,000 images with annotations for various anatomical structures and abnormalities. To adapt Chest ImaGenome to our settings (L, R, and C), we use the anatomy annotation to identify abnormality labels for the left lung, right lung, and heart. We adhere to the data split provided by [61] for Chest ImaGenome and follow the procedure outlined in Section 5 for DiagnosedGaze++. Note that REFLACX and EGD are not evaluated because their samples and annotations come from MIMIC-CXR, the same source as Chest ImaGenome.

To train the heatmap prediction-based interpretable approaches on Chest ImaGenome, we initially train the heatmap predictor on DiagnosedGaze++'s training set. Subsequently, we predict the heatmaps for the left lung, right lung, and heart for each image in the Chest ImaGenome dataset. Following this, we train the classifier head using the images weighted by the predicted heatmaps, along with their corresponding classification labels. We ensure that the training set used in the initial phase does not overlap with the validation and test set of the Chest ImaGenome dataset.

6.2. Implementation details

The ViT adapter is an 8-layer vision transformer with dimensions of 240, 6 attention heads, and an input patch size of 16×16. The BiomedCLIP's visual encoder is a 12-layer ViT-B/16 pretrained on resolution 224^2. The BiomedCLIP's text encoder is a 12-layer BERT with a vocab size of 30,522. We freeze both the text encoder and visual encoder of BiomedCLIP in the attention prediction and classification stages. The Convolutional layer in the fusion block has 1×1 convolution kernel, stride of 1, and 240 channels. The Linear layer in the fusion block has 512 as the input dimension and 240 as the output dimension. The MLPs in the Intensity Decoder have 3 fully connected layers and a hidden dimension of 256. We proceed to train them with a learning rate of 0.0001, batch size of 16, 60,000 iterations, and AdamW optimizer [64]. The training process takes roughly 4 h on a single Quadro RTX 8000 GPU.

6.3. Metrics

To quantify the performance at capturing the radiologist's intensity, we use the mean of L_2 (mL2), L_1 (mL1), and peak signal-to-noise ratio (mPSNR) over all samples. On the other hand, we also need to measure how well the attention can filter out irrelevant pixels by using intersection over union on foreground (fgIoU), background (bgIoU), and frequency-weighted IoU (fwIoU). To measure how good the classifiers are, we use the standard metrics Accuracy, AUC, and F1.

Intensity metrics. Given the predicted gaze attention $\hat{a}$ and the ground truth gaze attention a of a sample, we compute L_2, L_1, and PSNR by:

$$L_2 = \|\hat{a} - a\|_2 = \sqrt{\sum_{i=1,j=1}^{H,W} (\hat{a}_{i,j} - a_{i,j})^2}, \tag{15}$$

$$L_1 = \|\hat{a} - a\|_1 = \sum_{i=1,j=1}^{H,W} |\hat{a}_{i,j} - a_{i,j}|, \tag{16}$$

$$\text{PSNR} = 10 \log_{10}\left(\frac{MAX_I^2}{MSE(\hat{a}, a)}\right) = 10 \log_{10}\left(\frac{W+H}{L_2^2}\right), \tag{17}$$

where H is the height, W is the width, $a_{i,j}$ is the true attention value of pixel (i, j), and $\hat{a}_{i,j}$ is the predicted attention value, MAX_I is the maximum possible intensity value (in our case it is 1), and $MSE = \frac{1}{W+H}\sum_{i=1,j=1}^{H,W}(\hat{a}_{i,j} - a_{i,j})^2 = \frac{L_2^2}{W+H}$. The reported values of mL2, mL1, and mPSNR are the means of L_2, L_1, and PSNR across all samples, respectively.

Location metrics. Given the predicted gaze mask $\hat{m}$ and the ground truth gaze mask m, we compute fgIoU, bgIoU, and fwIoU by:

$$\text{fgIoU} = \frac{|\hat{m}_{fg} \cap m_{fg}|}{|\hat{m}_{fg} \cup m_{fg}|}, \tag{18}$$

$$\text{bgIoU} = \frac{|\hat{m}_{bg} \cap m_{bg}|}{|\hat{m}_{bg} \cup m_{bg}|}, \tag{19}$$

$$\text{fwIoU} = \frac{1}{f_{fg} + f_{bg}}(f_{fg}\,\text{fgIoU} + f_{bg}\,\text{bgIoU}), \tag{20}$$

where $\hat{m}_{fg}$ is the set of pixels that represents the foreground, in our case, where the pixel value equals 1, and similarly we obtain m_{fg}. For the set of pixels whose value equals 0, we obtain the predicted background mask $\hat{m}_{bg}$ and ground truth background mask m_{bg}. f_{fg} and f_{bg} are the frequencies of the foreground pixels and background pixels, respectively. The frequency is defined as all occurrences in a sample of a value, i.e. 0 for background, and 1 for foreground. The reported values of fgIoU, bgIoU, and fwIoU are averaged across the whole dataset.

Diagnosis metrics. We use the standard Accuracy, AUC, and F1 to measure the performance of classifiers on diagnosing.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i), \tag{21}$$

where N is the total number of samples, $\hat{y}_i$ is the predicted finding for sample i, y_i is the ground truth finding for sample i, $\mathbb{1}(\cdot)$ is the indicator function, which equals 1 if the condition inside the parentheses is true, and 0 otherwise.

AUC is calculated by plotting the true positive rate, i.e. $TPR = \frac{TP}{TP+FN}$, against the false positive rate, i.e. $FPR = \frac{FP}{FP+TN}$, where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, TN is the number of true negatives, at various classification thresholds and then computing the area under the resulting curve.

Finally, we compute F1 Score

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{22}$$

where $\text{precision} = \frac{TP}{TP+FP}$ and $\text{recall} = \frac{TP}{TP+FN}$.

7. Experiment results

7.1. Comparison

Methods. To evaluate the performance of our ItpCtrl-AI, we compare our model under various aspects:

- Black-box approach: In terms of Locaiton and Intensity, we train a ResNet-101 [2] on our settings. Then we use Relevance-CAM [11], Grad-CAM [12], Grad-CAM++ [17], and Integrated Grad-CAM [13] to extract the attention heatmaps. In term of Decision, we train Karargyris et al.'s method (EfficientNet backbone) [19], Medical MAE [65], and TA-DCL [66]. On the Chest ImaGenome classification task, we do not have gaze supervision for Karargyris et al. [19], so we remove their heatmap decoder head for evaluation.
- Prototype-based approaches: We deploy PIPNet [41], XProtoNet [67], ProtoPool [68], and ProtoPNet [14].
- Replacing attention predictors: we substitute our adapter with TransUNet [19,32] or PnP-AdaNet [33]. We train these alternative segmentation methods using our attention and pseudo-groundtruth masks. Subsequently, we apply the same masking technique as in ItpCtrl-AI before classification.

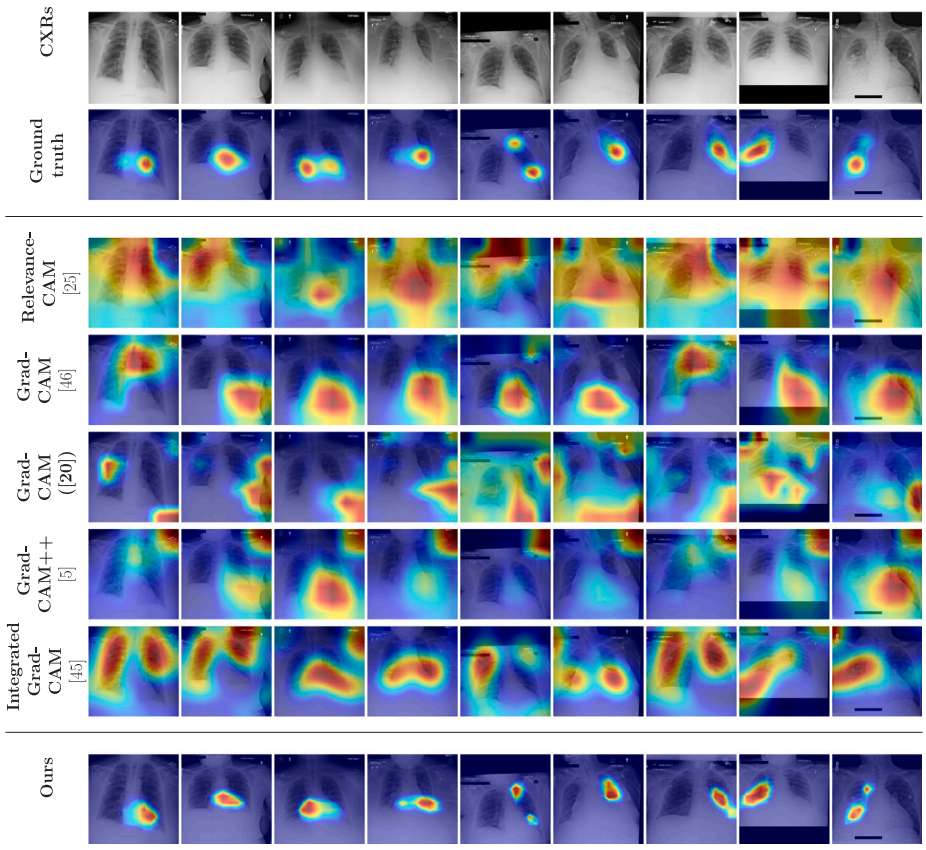

Fig. 5. Comparison of the results between various *CAM methods* and ours ItpCtrl-AI.

Note that, all methods, except ItpCtrl-AI, are trained on separate subsets, i.e. C, L, and R subsets, because we need to use one input to infer on three anatomies. Then, we take the average of performance on all subsets to get the final scores. Meanwhile, ItpCtrl-AI is trained on only the M subset. We evaluate all methods 5 times with cross-validation and reported the mean and dispersion scores for all metrics.

7.2. Qualitative results

Figs. 5–7 show the difference between our results and other results from CAM, segmentation, and prototype-based methods. Our approach yields notably more accurate attention, attributed to an architecture specifically designed with a focus on replicating radiologist-based attention.

Despite training across three distinct subsets, CAM methods result in inaccurate and unreliable attention heatmaps due to the lack of specific constraints, as shown in Fig. 5. Note that, although [19]'s predicted attention heatmap is not too far off and its accuracy is 73.72%(Table 7), its Grad-CAM visualizations (Fig. 5) show that [19] use mostly irrelevant information to classify and produce attention heatmaps.

On the other hand, Fig. 6 highlights the capability of segmentation methods to identify relevant areas effectively. Nonetheless, our methodology aligns more closely with the ground truth, largely due to the L_2 attention loss. This loss is instrumental in refining the model's attention predictions, considering that the ground truth is derived from frequency counts. TransUNet, PnP-AdaNet, and the method described in [19] utilize solely segmentation loss, such as binary cross-entropy loss, designed to manage probability values. Although the attention heatmap values range from 0 to 1 for both our method and these segmentation approaches, the underlying characteristics of these values diverge, leading to differences in performance outcomes. This effect is further validated in our ablation study (Section 8.2).

Regarding prototype-based approaches, Fig. 7 shows that state-of-the-art methods, even when trained without gaze supervision, can achieve commendable localization results qualitatively, particularly XProtoNet. ProtoPNet, being one of the earlier prototype-based methods, often struggles to localize important areas. XProtoNet, on the other hand, performs well in localizing areas of interest. However, the inclusion of gaze supervision provides ItpCtrl-AI with an overall performance boost, especially noticeable in intensity metrics.

7.3. Quantitative results

Location and Intensity. As shown in Table 6, our method achieves superior performance over other methods. Among the CAM-based methods, Integrated Grad-CAM has the highest scoring, but its scores are still lower than those of the methods that directly predict attention heatmaps. For example, Integrated Grad-CAM has a fgIoU score lower than ours by approximately 25 units. In terms of "where to look at", [19]'s IoU scores closely match our method. In particular, ours has a slightly increase fgIoU than [19] by 0.23, and our method has a better bgIoU by 9.33 unit. Regarding prototype-based methods, most can decently localize the left lung, right lung, and heart, particularly XProtoNet. The best prototype-based method evaluated on our dataset, XProtoNet, achieves scores comparable to heatmap prediction methods in location metrics. However, without gaze supervision, these methods fail to achieve better intensity scores. Our method outperforms existing approaches in both location and intensity metrics, demonstrating superior performance in predicting radiologists' gaze patterns on chest X-rays compared to the current SOTAs.

Diagnosis. Tables 7 and 8 show that our method achieves the highest scores across all metrics. For example on AUC, ItpCtrl-AI achieves

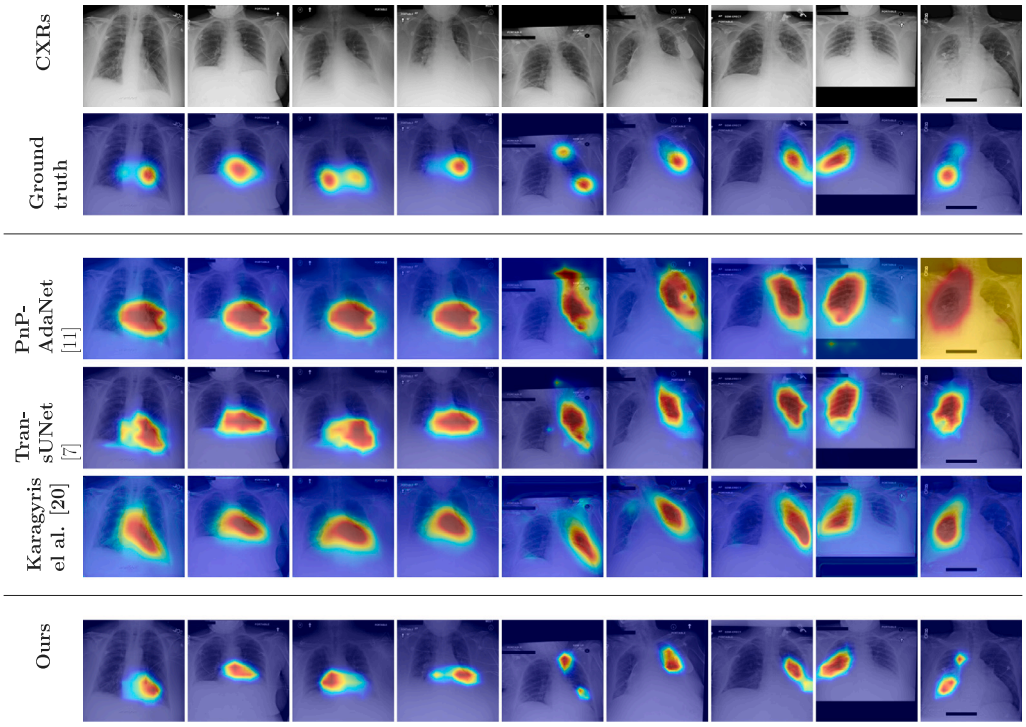

Fig. 6. Comparison of the results between various *segmentation methods* and our ItpCtrl-AI.

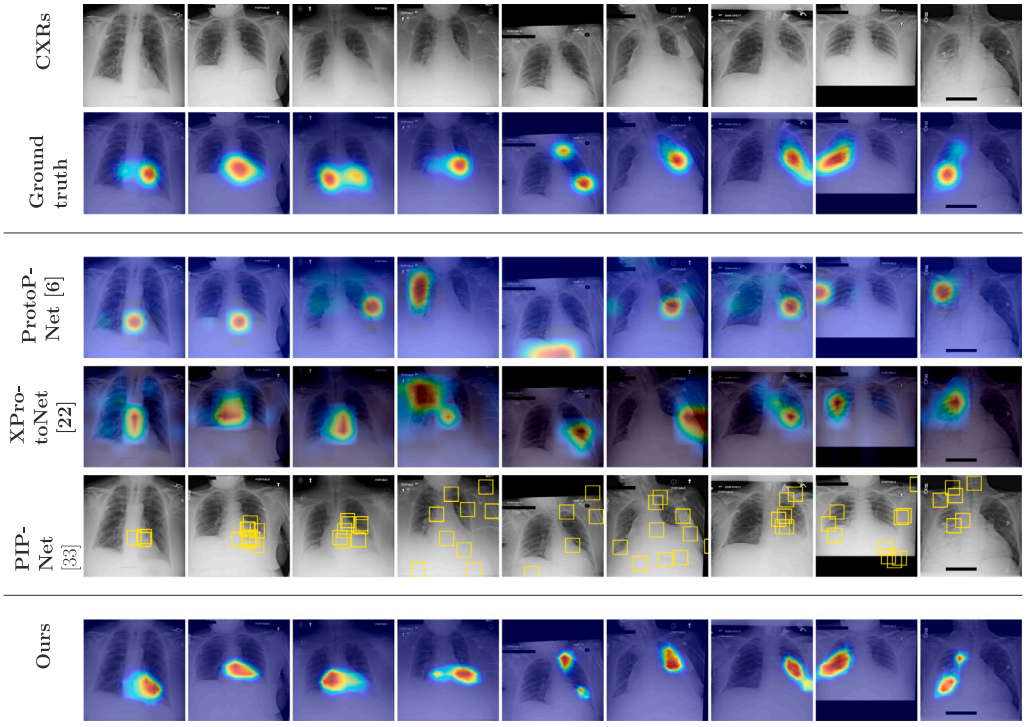

Fig. 7. Comparison of the results between various *prototype-based methods* and our ItpCtrl-AI. For local explanation, PIPNet uses patches (yellow boxes indicate the chosen patches), XProtoNet uses occurrence maps to highlight relevant areas, and ProtoPNet employs a similarity matrix for localization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

76.81% and 88.76%, respectively, even though ItpCtrl-AI utilizes only a portion of the input image defined by our predicted attention heatmap, while the other methods have access to the full image. Most methods perform well as classifiers in our settings, likely because our settings of C, L, and R are generally straightforward. For instance, Medical MAE achieves nearly 75% in AUC on DiagnosedGaze++, and EfficientNet (Karagyris et al. [19]) achieves an AUC of 86.53%. Notably, XProtoNet is effective in both location and intensity metrics and performs well in the classification task. Overall, our method performs well and shows strong results on both DiagnosedGaze++ and the public Chest ImaGenome dataset.

Table 6
The performance in Radiologists' Attention Modeling challenge in comparison with state-of-the-art methods. Our method stands out for its fine-grained localization and precision. Note that the Grad-CAM [19] method extracts the attention from [19] using Grad-CAM. CAM-based techniques derive heatmaps from black-box models. Prototype-based methods utilize local explanations specific to each model, such as patches in PIPNet or occurrence maps in XProtoNet. Heatmap prediction methods are directly supervised using gaze heatmap ground truth data.

Attention type	Methods	Location			Intensity		
		fgIoU↑	bgIoU↑	fwIoU↑	mPSNR↑	mL1↓	mL2↓
CAM-based	Relevance-CAM [11]	33.12 ± 1.02	41.72 ± 0.50	40.85 ± 1.36	4.38 ± 0.29	0.56 ± 0.03	0.30 ± 0.02
	Grad-CAM [12]	36.78 ± 0.92	79.55 ± 0.75	75.78 ± 1.25	9.14 ± 0.35	0.22 ± 0.03	0.15 ± 0.04
	Grad-CAM++ [17]	27.50 ± 0.88	81.02 ± 1.10	76.77 ± 1.11	9.65 ± 0.41	0.21 ± 0.01	0.09 ± 0.02
	Integrated Grad-CAM [13]	30.33 ± 1.13	83.90 ± 0.90	79.44 ± 1.38	11.60 ± 0.52	0.17 ± 0.02	0.08 ± 0.01
	Grad-CAM (from [19])	24.73 ± 0.74	56.22 ± 1.20	53.47 ± 0.95	8.50 ± 0.40	0.27 ± 0.04	0.13 ± 0.03
Prototype-based	ProtoPNet [14]	41.78 ± 2.03	85.08 ± 1.14	81.23 ± 1.25	10.46 ± 0.74	0.19 ± 0.04	0.09 ± 0.02
	ProtoPool [68]	40.34 ± 1.13	85.44 ± 0.91	81.47 ± 0.23	9.56 ± 0.11	0.20 ± 0.03	0.10 ± 0.01
	PIPNet [41]	43.46 ± 1.43	86.27 ± 0.45	82.43 ± 0.31	10.25 ± 0.40	0.18 ± 0.02	0.10 ± 0.02
	XProtoNet [67]	49.53 ± 0.66	89.36 ± 0.99	85.66 ± 0.59	12.33 ± 0.28	0.12 ± 0.02	0.08 ± 0.01
Heatmap Prediction	Karargyris et al. [19]	55.67 ± 0.65	84.88 ± 0.92	82.10 ± 1.08	12.50 ± 0.60	0.18 ± 0.03	0.09 ± 0.02
	TransUNet [32]	51.65 ± 1.02	91.60 ± 1.15	87.95 ± 0.95	12.20 ± 0.75	0.10 ± 0.03	0.07 ± 0.03
	PnP-AdaNet [33]	48.60 ± 0.87	90.85 ± 1.05	86.75 ± 1.00	11.40 ± 0.58	0.09 ± 0.02	0.06 ± 0.01
	ItpCtrl-AI (Ours)	**55.90 ± 0.60**	**94.21 ± 0.55**	**90.45 ± 0.70**	**17.65 ± 0.75**	**0.06 ± 0.02**	**0.03 ± 0.02**

Table 7
The performance in Decision-making challenge for all classifiers on DiagnosedGaze++. Black-box methods directly predict labels from inputs without explicit visual explanations. Prototype-based approaches learn to explain their predictions automatically during end-to-end training, without gaze heatmap supervision. Heatmap Prediction methods first generate a heatmap, then use it to inform diagnosis. Following this categorization, Karagyris et al. [19]'s work is classified as a black-box method, as its predicted heatmap stems from the same bottleneck feature used for classification, rather than directly informing the diagnostic process.

Model type	Model	Accuracy(%)	F1(%)	AUC(%)
Black-box	Resnet-101 [2]	70.20 ± 0.21	68.12 ± 0.15	70.91 ± 0.13
	TA-DCL [66]	64.10 ± 0.92	63.54 ± 0.88	67.18 ± 0.51
	Medical MAE [65]	73.46 ± 0.42	75.60 ± 0.68	74.92 ± 0.25
	Karargyris et al. [19]	73.72 ± 0.33	73.44 ± 0.10	74.53 ± 0.03
Prototype-based	PIPNet [41]	73.10 ± 0.45	75.27 ± 0.45	74.77 ± 0.45
	ProtoPool [68]	70.91 ± 1.01	71.67 ± 0.50	72.03 ± 0.12
	XProtoNet [67]	71.41 ± 0.12	72.21 ± 0.41	74.90 ± 0.74
	ProtoPNet [14]	63.10 ± 0.94	60.27 ± 0.62	62.56 ± 0.31
Heatmap prediction	TransUNet [32]	72.33 ± 1.16	72.16 ± 0.25	72.71 ± 0.27
	PnP-AdaNet [33]	73.65 ± 1.04	71.53 ± 0.22	73.18 ± 0.21
	ItpCtrl-AI (Ours)	**75.66 ± 0.27**	**76.51 ± 0.20**	**76.81 ± 0.35**

Table 8
The performance in Decision-making challenge for all classifiers on the Chest ImaGenome dataset. Black-box methods directly predict labels from inputs without explicit visual explanations. Prototype-based approaches learn to explain their predictions automatically during end-to-end training, without gaze heatmap supervision. Heatmap Prediction methods first generate a heatmap, then use it to inform diagnosis.

Model type	Model	Accuracy(%)	F1(%)	AUC(%)
Black-box	Resnet-101 [2]	76.41 ± 0.85	75.62 ± 0.91	86.29 ± 1.12
	TA-DCL [66]	69.78 ± 1.23	70.94 ± 1.15	75.81 ± 1.09
	Medical MAE [65]	73.69 ± 0.94	77.52 ± 1.01	83.12 ± 0.87
	Karargysis et al. [19](EfficientNet [69])	76.97 ± 0.89	75.19 ± 1.02	86.53 ± 0.97
Prototype-based	PIPNet [41]	75.86 ± 1.14	73.48 ± 1.09	77.14 ± 1.21
	ProtoPool [68]	67.96 ± 1.18	67.38 ± 1.07	71.09 ± 1.13
	XProtoNet [67]	77.84 ± 0.87	78.94 ± 0.78	87.88 ± 0.64
	ProtoPNet [14]	75.32 ± 1.02	76.89 ± 0.84	84.57 ± 1.18
Heatmap prediction	TransUNet [32]	77.98 ± 0.93	78.24 ± 0.72	86.75 ± 0.88
	PnP-AdaNet [33]	77.69 ± 0.76	78.52 ± 0.95	85.71 ± 1.05
	ItpCtrl-AI (Ours)	80.39 ± 0.73	80.41 ± 0.82	88.76 ± 0.69

8. Ablation study

8.1. Heatmap prediction of a particular setting

The use of querying to predict findings on a particular region can make the model good at some findings, while bad at others. To demonstrate that our model can overcome this limitation, we use the same checkpoint trained on the final dataset (M setting), and then we test on separate settings: C, L, and R. As shown in Table 9, our model achieves high performance with a marginal difference between the full setting (M) versus subsetting C, L, and R. Table 9 also shows the robustness of our model across all settings.

Table 9
Ablation study: Attention prediction of particular settings.

Settings	fgIoU↑	fwIoU↑	mPSNR↑	mL1↓
C	54.22 ± 0.15	90.68 ± 0.92	17.45 ± 0.38	0.06 ± 0.01
L	54.75 ± 1.28	89.88 ± 1.05	17.51 ± 1.02	0.07 ± 0.02
R	55.58 ± 0.33	90.41 ± 0.88	16.09 ± 0.25	0.09 ± 0.01
M	55.90 ± 0.60	90.45 ± 0.70	17.65 ± 0.75	0.06 ± 0.02

Table 10
Ablation study: the impact of attention and mask losses.

Losses		Location		Intensity	
L_h	L_s	fgIoU↑	fwIoU↑	mPSNR↑	mL2↓
✓	✗	21.54 ± 1.20	84.04 ± 1.15	15.58 ± 1.30	0.03 ± 0.01
✗	✓	54.67 ± 1.25	85.87 ± 0.95	12.34 ± 1.10	0.05 ± 0.02
✓	✓	**55.90 ± 0.60**	**90.45 ± 0.70**	**17.65 ± 0.75**	**0.03 ± 0.02**

8.2. The importance of losses

Both losses are equally important. Without the gaze mask loss L_m, the gradient flow of L_h, which is based on L_2, is not enough for the model to learn where to look, and it can easily collapse to a local minima where a metric like mL2 is good, but other metrics like fgIoU are bad. As shown in Table 10, fgIoU dramatically drops to 21.54, while mL2 is 0.03. On the other hand, using only the L_m loss can only make the model predict general area instead of gaze attention, as shown in Section 7.2, and gives suboptimal results in mPSNR with 12.34 points. Therefore, we use masks created from the attention heatmaps together with cross-entropy loss and dice loss to guide the model as the gaze mask loss L_m, and the L_2 loss as heatmap loss L_h, which is close to the nature of attention heatmap (based on counting as described in Section 5).

8.3. The importance of anatomic-driven adapter

The Anatomic-driven Adapter (ADA) plays a vital role, as evidenced by the fact that without it, the performance is only marginally better than the CAM black-box method. The ADA is essential because it effectively serves as the heatmap predictor. To emphasize the necessity of a deep adapter, we conduct an experiment where the adapter was replaced with an MLP that receives input from both the visual and text features extracted from BiomedCLIP. The visual feature was obtained from the last layer (12th layer) of BiomedCLIP's visual encoder, while the text feature was derived from the final layer of the text encoder, as in Section 4. The results presented in Table 11 show that ItpCtrl-AI without ADA only marginally outperforms the black-box approaches and achieves inferior results compared to the complete ItpCtrl-AI. Furthermore, the gaze attention predictability significantly decreases and yields an mPSNR lower than that of Integrated Grad-CAM (Table 6) when ADA is not employed. These findings emphasize the fact that without the ADA utilized in ItpCtrl-AI, the system is unable to effectively adapt to the gaze attention prediction task and, consequently, fails to provide adequate attention for the classifier.

8.4. The importance of scaling vector s

We define a learnable scaling vector s in Section 4 to help the model learn. From the output $f_{a_L} \in \mathbb{R}^{W\times H\times D}$, it is true that we can naively create the final output by taking the mean of the last dimension. However, as shown in Table 12, the inflexibility of naively averaging the feature space effectively prevents the model from learning.

8.5. Performance on various classifier backbones

The effectiveness of the classifier could be from the strong backbone of our classifier. To determine how strong of the effect of our chosen Visual Encoder, we change our Visual Intense Encoder in Section 4.4, which is BiomedCLIP's Visual Encoder, to Resnet-101 [2] and EfficientNet-b0 [69]. We also allow these encoders to learn, instead of using a frozen pretrained from other dataset. Table 13 presents the results that demonstrate the lack of significant impact on performance when altering the backbone, thus suggesting that the primary factor influencing outcomes lies in the masking of input prior to the classifier.

8.6. Weight initialization

When it comes to training neural networks, the effective initialization of network weights holds paramount importance in constructing robust and generalizable models. Usually, this objective is accomplished through the utilization of pre-trained weights. However, in our specific pipeline, this process is non-trivial. In Table 14, we compare the performance of using pretrained ImageNet weights versus random weights on our adapter.

- We utilize pretrained weights with eight transformer layers sourced from Huggingface's model zoo, trained on the ImageNet-21k dataset [3]. These weights are applied to our adapter in a one-to-one mapping, matching layer orders (e.g., 0 to 0, 1 to 1, 2 to 2), while the fusion block is initialized randomly.
- In the random initialization scenario, we use PyTorch's default initialization for all layers.

The results show no performance improvement, possibly due to the disparity and domain gap between domain of natural images in ImageNet and that of medical images (CXR in our case). While medical pretrained weights could be beneficial, developing an effective pretraining strategy for our adapter, which handles both textual and visual data, remains challenging. Current leading strategies [4,70] often employ separate encoders for text and vision, rather than a unified one. As a result, we choose random weight initialization for the adapter to maintain its versatility.

8.7. The importance of end-to-end training scheme

While the general pipeline is the same, different training schemes can bring different performance. In our findings, end-to-end training can help improve all components altogether. In particular, the results shown in Table 15 demonstrate that the end-to-end training approach slightly outperforms the two-stage training across most metrics. Specifically, it achieves marginally higher fgIoU, bgIoU, and fwIoU, indicating a more accurate localization of relevant areas. In intensity metrics, it also performs better, with higher mPSNR and lower mL1 and mL2, suggesting that it can more accurately capture the radiologists' attention. Furthermore, the end-to-end approach shows a slight improvement in diagnostic metrics, with higher accuracy, F1 score, and AUC, indicating a more effective overall diagnosis capability. These comparisons highlight the advantages of the end-to-end training scheme in enhancing model performance across location accuracy, intensity of focus, and diagnostic reliability.

8.8. Choosing layers for fusion

The information from BiomedCLIP plays a paramount role in our pipeline. Specifically, relying solely on our adapter leads to a meager fgIoU score of 27.39 (refer to Table 16, w/o. fusion setting).

Then, we follow [59] to determine the optimal fusion layer. By default, we use the feature of the last layer of BiomedCLIP's text encoder as our text feature f_t. Subsequently, we proceed to merge these text features with features from different layers of BiomedCLIP's visual encoder, labeled as f_{v_j} where $j \in \{0, 1, 2, \dots, 12\}$, alongside features in a different position in our adapter, referred f_{a_i} where $i \in \{0, 1, 2, \dots, 8\}$. This fusion process generates the combined feature f'_{a_i} (as described in Eq. (1)). Let 0th be the 0 layer. We run multiple fusion settings as below:

- w/o. fusion: there is no fusion operations happening. We train our adapter from scratch in this setting.
- 0: visual encoder's 0th layer + text feature + our adapter's 0th layer, i.e.

$$f'_{a_0} = f_{v_0} + f_t + f_{a_0}$$

Table 11
Ablation study: Impact of Anatomic-driven Adapter (ADA) on ItpCtrl-AI.

Weight initialization	Location		Intensity		Diagnosis	
	fgIoU↑	bgIoU↑	mPSNR↑	mL1↓	F1(%)↑	AUC(%)↑
w/o ADA	46.20 ± 1.15	89.90 ± 0.85	10.75 ± 0.92	0.09 ± 0.02	73.30 ± 1.20	72.15 ± 1.10
w/ ADA	**55.90 ± 0.60**	**94.21 ± 0.55**	**17.65 ± 0.75**	**0.06 ± 0.02**	**76.51 ± 0.20**	**76.81 ± 0.35**

Table 12
Ablation study: Impact of the scaling vector s on anatomic-driven adapter.

Settings	fgIoU↑	fwIoU↑	mPSNR↑	mL1↓
w/o s	17.81 ± 0.12	66.03 ± 0.20	11.62 ± 0.08	0.15 ± 0.01
w/ s	**55.90 ± 0.60**	**90.45 ± 0.70**	**17.65 ± 0.75**	**0.06 ± 0.02**

Table 13
Ablation study: Our ItpCtrl-AI model on various classifier backbones.

Model	Accuracy(%)	F1(%)	AUC(%)
Ours (Resnet-101)	74.35 ± 0.15	74.46 ± 0.12	74.40 ± 0.14
Ours (EfficientNet-b0)	75.39 ± 0.11	75.16 ± 0.10	75.43 ± 0.09
Ours	**75.66 ± 0.27**	**76.51 ± 0.20**	**76.81 ± 0.35**

- 3rd: visual encoder's 3rd layer + text feature + our adapter's 0th layer, i.e.

$$f'_{a_0} = f_{v_3} + f_t + f_{a_0}$$

- 6th: visual encoder's 6th layer + text feature + our adapter's 0th layer, i.e.

$$f'_{a_0} = f_{v_6} + f_t + f_{a_0}$$

- 9th: visual encoder's 9th layer + text feature + our adapter's 0th layer, i.e.

$$f'_{a_0} = f_{v_9} + f_t + f_{a_0}$$

- 12th: visual encoder's 12th layer + text feature + our adapter's 0th layer, i.e.

$$f'_{a_0} = f_{v_{12}} + f_t + f_{a_0}$$

- {9,12}th: visual encoder's {9,12}th layers + text feature + our adapter's{0,1}st layers, i.e.

$$f'_{a_0} = f_{v_9} + f_t + f_{a_0}$$
$$f'_{a_1} = f_{v_{12}} + f_t + f_{a_1}$$

- {6,9}th: visual encoder's {6,9}th layers + text feature + our adapter's {0,1}st layers, i.e.

$$f'_{a_0} = f_{v_6} + f_t + f_{a_0}$$
$$f'_{a_1} = f_{v_9} + f_t + f_{a_1}$$

- {3,6,9}th: visual encoder's {3,6,9}th layers + text feature + our adapter's {0,1,2}th layers, i.e.

$$f'_{a_0} = f_{v_3} + f_t + f_{a_0}$$
$$f'_{a_1} = f_{v_6} + f_t + f_{a_1}$$
$$f'_{a_2} = f_{v_9} + f_t + f_{a_2}$$

- {0,3,6,9}th: visual encoder's {0,3,6,9}th layers + text feature + our adapter's{0,1,2,3}th layers, i.e.

$$f'_{a_0} = f_{v_0} + f_t + f_{a_0}$$
$$f'_{a_1} = f_{v_3} + f_t + f_{a_1}$$
$$f'_{a_2} = f_{v_6} + f_t + f_{a_2}$$
$$f'_{a_3} = f_{v_9} + f_t + f_{a_3}$$

- {0,3,6,9,12}th: visual encoder's {0,3,6,9,12}th layers + text feature + our adapter's {0,1,2,3,4}th layers, i.e.

$$f'_{a_0} = f_{v_0} + f_t + f_{a_0}$$
$$f'_{a_1} = f_{v_3} + f_t + f_{a_1}$$
$$f'_{a_2} = f_{v_6} + f_t + f_{a_2}$$
$$f'_{a_3} = f_{v_9} + f_t + f_{a_3}$$
$$f'_{a_4} = f_{v_{12}} + f_t + f_{a_4}$$

The findings in Table 16 demonstrate a substantial performance enhancement when fusing features from deeper layers, such as 9 and 12, within a single fusion configuration. In pursuit of further improving results, we extensively explore various combinations of positions for fusing features in both the BiomedCLIP visual encoder and our adapter. We discover that the {0, 3, 6, 9}th positions, which is the proposed setting in Section 4, yield the most superior outcomes.

9. Limitations

Our study provides promising insights into radiological image interpretation; however, we acknowledge certain limitations that we aim to address in future work.

Our framework utilizes pre-trained VLMs (BiomedCLIP) for extracting features. BiomedCLIP was trained on PMC-15M dataset containing 15 million biomedical image-text pairs collected from 4.4 million scientific articles. While BiomedCLIP is recognized one of the SOTA pre-trained models for medical imaging, it is a generalist 2D medical imaging model and is not specifically designed for CXR. Therefore, a more specialized model tailored to represent CXR data would be beneficial. To address this limitation, it is essential to develop a specialized CXR model in the future, which necessitates access to a large-scale CXR dataset. We believe one promising approach to achieve this is through self-supervised learning, which could leverage the vast amount of unlabeled CXR data to create a more tailored and robust representation of chest radiographs.

While CXR images are widely available, quick to perform, and relatively inexpensive compared to CT scans, they may not reveal subtle or small pathologies such as small tumors, early-stage lung diseases, or subtle fractures. Thus, it is necessary to extend the current developing computational model to model radiologists' intentions on CT scans. However, obtaining gaze sequence data on CT images poses significant challenges. We have taken this challenge into consideration and our clinical team and scientific team have worked together to obtain such data. Developing an end-to-end interpretable and controllable AI on CT scans is our ongoing research and we are at the stage of CT scans acquisition and preprocessing. It is important to emphasize that working with CT scans presents substantially greater challenges compared to CXR images, due to their three-dimensional nature, increased complexity, and the vast amount of data each scan contains. This complexity not

Table 14
Our pipeline performance under different weight initialization.

Weight initialization	Location		Intensity		Diagnosis	
	fgIoU↑	bgIoU↑	mPSNR↑	mL1↓	F1(%)↑	AUC(%)↑
ImageNet init.	48.95 ± 0.20	92.49 ± 0.25	16.11 ± 0.12	0.11 ± 0.01	74.43 ± 0.15	75.34 ± 0.18
Random init.	**55.90 ± 0.60**	**94.21 ± 0.55**	**17.65 ± 0.75**	**0.06 ± 0.02**	**76.51 ± 0.20**	**76.81 ± 0.35**

Table 15
Ablation study: Our pipeline performance in comparison with two-stage training.

Training scheme	Location		Intensity		Diagnosis	
	fgIoU↑	bgIoU↑	mPSNR↑	mL1↓	F1(%)↑	AUC(%)↑
Two-stage	55.83 ± 0.18	93.28 ± 0.22	16.90 ± 0.15	0.08 ± 0.01	76.24 ± 0.14	76.18 ± 0.13
Ours	**55.90 ± 0.60**	**94.21 ± 0.55**	**17.65 ± 0.75**	**0.06 ± 0.02**	**76.51 ± 0.20**	**76.81 ± 0.35**

Table 16
Ablation study: our method with different layer positions for the fusion process.

Fusion layers	Location		Intensity		Diagnosis	
	fgIoU↑	bgIoU↑	mPSNR↑	mL1↓	F1(%)↑	AUC(%)↑
w/o. fusion	27.39 ± 0.14	79.95 ± 0.20	9.15 ± 0.12	0.17 ± 0.03	69.82 ± 0.15	69.67 ± 0.14
0th	28.62 ± 0.13	80.58 ± 0.18	10.80 ± 0.10	0.15 ± 0.02	70.95 ± 0.12	70.66 ± 0.13
3rd	43.62 ± 0.22	86.10 ± 0.25	13.36 ± 0.17	0.14 ± 0.01	74.71 ± 0.18	74.39 ± 0.15
6th	44.80 ± 0.20	89.12 ± 0.30	15.90 ± 0.20	0.11 ± 0.03	76.01 ± 0.16	75.37 ± 0.14
9th	51.48 ± 0.18	90.71 ± 0.22	16.12 ± 0.15	0.10 ± 0.02	76.30 ± 0.14	75.88 ± 0.12
12th	48.29 ± 0.25	87.10 ± 0.20	15.87 ± 0.14	0.09 ± 0.01	77.47 ± 0.18	75.74 ± 0.15
{9,12}th	49.10 ± 0.18	88.95 ± 0.24	16.69 ± 0.16	0.09 ± 0.02	77.12 ± 0.15	76.08 ± 0.13
{6,9}th	51.06 ± 0.14	91.23 ± 0.25	17.12 ± 0.18	0.09 ± 0.02	74.96 ± 0.13	75.96 ± 0.12
{3,6,9}th	54.86 ± 0.22	93.09 ± 0.28	17.37 ± 0.17	0.08 ± 0.02	76.30 ± 0.15	75.94 ± 0.14
{0,3,6,9}th	**55.90 ± 0.60**	**94.21 ± 0.55**	**17.65 ± 0.75**	**0.06 ± 0.02**	**76.51 ± 0.20**	**76.81 ± 0.35**
{0,3,6,9,12}th	54.93 ± 0.24	91.71 ± 0.28	16.84 ± 0.20	0.07 ± 0.01	76.50 ± 0.18	75.32 ± 0.14

only affects data collection and preprocessing but also significantly increases the computational requirements and model complexity needed to effectively analyze and interpret CT scans.

Despite some limitations mentioned earlier, our work is one of the first attempts to model radiologists' intentions, aiming to create a controllable and interpretable decision-making AI framework. We hope that our findings will inspire future research in the large area of trustworthy AI in healthcare.

10. Discussion and conclusion

In summary, we have successfully developed a unified controllable, and interpretable pipeline that innovatively addresses the challenges in medical imaging diagnostics. Our approach stands out for its ability to generate anatomical attention heatmaps and predict abnormal findings in chest X-ray images. By leveraging anatomical prompts, our model offers a unique layer of control and flexibility, allowing users to specify the type of heatmap and diagnosis they seek. Moreover, the framework enhances interpretability by mirroring the focus and intensity of radiologists' gaze during their diagnostic process. This not only makes the model more reliable but also demystifies the reasoning behind medical diagnoses, making a significant move away from traditional black-box approaches.

Additionally, we have introduced a semi-automatic filtering process to produce a high-quality gaze dataset that provides gaze attention heatmaps, masks, and abnormality annotation. We hope the release of this dataset will advance the field, especially in helping future efforts to make CXR abnormality classification more interpretable.

From a broader perspective, our work is novel in its attempt to reverse engineer the diagnostic focus of radiologists. This is a pivotal step toward improving the transparency and reliability of AI systems in medical imaging.

CRediT authorship contribution statement

Trong-Thang Pham: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jacob Brecheisen:** Data curation. **Carol C. Wu:** Writing – review & editing, Validation. **Hien Nguyen:** Writing – review & editing, Project administration, Investigation. **Zhigang Deng:** Writing – review & editing. **Donald Adjeroh:** Writing – review & editing, Conceptualization. **Gianfranco Doretto:** Writing – review & editing. **Arabinda Choudhary:** Writing – review & editing. **Ngan Le:** Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Trong-Thang Pham reports financial support was provided by National Science Foundation. Trong-Thang Pham reports financial support was provided by National Institutes of Health. Carol Wu reports financial support was provided by National Institutes of Health. Carol Wu reports financial support was provided by National Science Foundation. Jacob Brecheisen reports financial support was provided by National Science Foundation. Hien Nguyen reports financial support was provided by National Science Foundation. Hien Nguyen reports financial support was provided by National Institutes of Health. Zhigang Deng reports financial support was provided by National Science Foundation. Zhigang Deng reports financial support was provided by National Institutes of Health. Ngan Le reports financial support was provided by National Science Foundation. Ngan Le reports financial support was provided by National Institutes of Health. Donald Adjeroh reports financial support was provided by National Science Foundation. Gianfranco Doretto reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Award No OIA-1946391, NSF 2223793 EFRI BRAID, National Institutes of Health (NIH) 1R01CA277739-01.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2024.103054.

References

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv preprint arXiv:2010.11929.
[2] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.
[3] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–55.
[4] Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. 2023, arXiv preprint arXiv:2303.00915.
[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[6] Muhammad K, Ullah A, Lloret J, Del Ser J, de Albuquerque VHC. Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Trans Intell Transp Syst 2020;22(7):4316–36.
[7] Al-Qizwini M, Barjasteh I, Al-Qassab H, Radha H. Deep learning algorithm for autonomous driving using googlenet. In: 2017 IEEE intelligent vehicles symposium. IEEE; 2017, p. 89–96.
[8] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.
[9] Fourcade A, Khonsari RH. Deep learning in medical image analysis: A third eye for doctors. J Stomatol Oral Maxillofac Surg 2019;120(4):279–88.
[10] Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. Jama 2018;320(11):1107–8.
[11] Lee JR, Kim S, Park I, Eo T, Hwang D. Relevance-cam: Your model already knows where to look. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 14944–53.
[12] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 618–26.
[13] Sattarzadeh S, Sudhakar M, Plataniotis KN, Jang J, Jeong Y, Kim H. Integrated grad-CAM: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing. IEEE; 2021, p. 1775–9.
[14] Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. Adv Neural Inf Process Syst 2019;32.
[15] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. 2017, arXiv preprint arXiv:1711.05225.
[16] van Sonsbeek T, Zhen X, Mahapatra D, Worring M. Probabilistic integration of object level annotations in chest X-ray classification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023, p. 3630–40.
[17] Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision. IEEE; 2018, p. 839–47.
[18] Rozenberg E, Freedman D, Bronstein A. Localization with limited annotation for chest X-rays. In: Machine learning for health workshop. PMLR; 2020, p. 52–65.
[19] Karargyris A, Kashyap S, Lourentzou I, Wu JT, Sharma A, Tong M, et al. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. Sci Data 2021.
[20] Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31.
[21] Pham TT, Ho N-V, Bui N-T, Phan T, Brijesh P, Adjeroh D, Doretto G, Nguyen A, Wu CC, Nguyen H, et al. FG-CXR: A Radiologist-Aligned Gaze Dataset for Enhancing Interpretability in Chest X-Ray Report Generation. In: Proceedings of the Asian Conference on Computer Vision. 2024, p. 941–58.
[22] Pham TT, Nguyen T-P, Ikebe Y, Awasthi A, Deng Z, Wu CC, Nguyen H, Le N. Gazesearch: radiology findings search benchmark. 2024, arXiv preprint arXiv:2411.05780.
[23] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 2921–9.
[24] Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020, p. 24–5.
[25] Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. Radiographics 2018;38(1):236–47.
[26] Król M, Król M. Learning from peers' eye movements in the absence of expert guidance: A proof of concept using laboratory stock trading, eye tracking, and machine learning. Cogn Sci 2019;43(2):e12716.
[27] Liu J, Zhao G, Fei Y, Zhang M, Wang Y, Yu Y. Align, attend and locate: Chest X-ray diagnosis via contrast induced attention network with limited supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 10632–41.
[28] Zhang Y, Shi Z, Wang H, Cui S, Zhang L, Wang L, et al. Automatic segmentation of lumbar Vertebra Anatomical Region based on hybrid swin-transformer network. In: 2023 IEEE 20th international symposium on biomedical imaging. IEEE; 2023, p. 1–5.
[29] You C, Dai W, Min Y, Staib L, Sekhon J, Duncan JS. Action++: Improving semi-supervised medical image segmentation with adaptive anatomical contrast. 2023, arXiv preprint arXiv:2304.02689.
[30] Shusharina N, Krishnaswamy D, Kinahan P, Fedorov A. Integrating deep learning algorithm for the lung segmentation with body-part-specific anatomical classification of medical imaging data hosted by medical imaging and data resource center (MIDRC). In: Medical imaging 2023: imaging informatics for healthcare, research, and applications, vol. 12469, SPIE; 2023, p. 180–6.
[31] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015, p. 234–41.
[32] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. Transunet: Transformers make strong encoders for medical image segmentation. 2021, arXiv preprint arXiv:2102.04306.
[33] Dou Q, Ouyang C, Chen C, Chen H, Heng PA. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. 2018, arXiv preprint arXiv:1804.10916.
[34] Lanfredi RB, Zhang M, Auffermann W, Chan J, Duong PA, Srikumar V, et al. REFLACX: Reports and eye-tracking data for localization of abnormalities in chest X-rays. PhysioNet; 2021.
[35] Caruana R. Multitask learning. Mach Learn 1997;28:41–75.
[36] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Interpretable machine learning: definitions, methods, and applications. 2019, arXiv preprint arXiv:1901.04592.
[37] Pham TT, Brecheisen J, Nguyen A, Nguyen H, Le N. I-AI: A controllable & interpretable AI system for decoding radiologists' intense focus for accurate CXR diagnoses. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024, p. 7850–9.
[38] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Mach Intell 2019;1(5):206–15.
[39] Nauta M, Van Bree R, Seifert C. Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 14933–43.
[40] Xue M, Huang Q, Zhang H, Cheng L, Song J, Wu M, et al. Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. 2022, arXiv preprint arXiv:2208.10431.
[41] Nauta M, Schlötterer J, van Keulen M, Seifert C. PIP-net: Patch-based intuitive prototypes for interpretable image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 2744–53.
[42] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. PMLR; 2018, p. 2668–77.
[43] Han Y, Chen C, Tewfik A, Glicksberg B, Ding Y, Peng Y, et al. Knowledge-augmented contrastive learning for abnormality classification and localization in chest X-rays with radiomics using a feedback loop. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022, p. 2465–74.
[44] Qi B, Zhao G, Wei X, Du C, Pan C, Yu Y, et al. GREN: graph-regularized embedding network for weakly-supervised disease localization in X-ray images. IEEE J Biomed Health Inf 2022;26(10):5142–53.
[45] Zhu H, Rohling R, Salcudean S. Multi-task unet: Jointly boosting saliency prediction and disease classification on chest X-ray images. 2022, arXiv preprint arXiv:2202.07118.

[46] Ouyang X, Karanam S, Wu Z, Chen T, Huo J, Zhou XS, et al. Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. IEEE Trans Med Imaging 2020;40(10):2698–710.
[47] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 4700–8.
[48] Yao L, Prosky J, Poblenz E, Covington B, Lyman K. Weakly supervised medical diagnosis and localization from multiple resolutions. 2018, arXiv preprint arXiv:1803.07703.
[49] Taslimi S, Taslimi S, Fathi N, Salehi M, Rohban MH. SwinCheX: Multi-label classification on chest X-ray images with transformers. 2022, arXiv preprint arXiv:2206.04246.
[50] Liu Q, Yu L, Luo L, Dou Q, Heng PA. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans Med Imaging 2020;39(11):3429–40.
[51] Liu F, Tian Y, Cordeiro FR, Belagiannis V, Reid I, Carneiro G. Self-supervised mean teacher for semi-supervised chest X-ray classification. In: Machine learning in medical imaging: 12th international workshop, MLMI 2021, held in conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, proceedings 12. Springer; 2021, p. 426–36.
[52] Gazda M, Plavka J, Gazda J, Drotar P. Self-supervised deep convolutional neural network for chest X-ray classification. IEEE Access 2021;9:151972–82.
[53] Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 3478–88.
[54] Wu J, Gur Y, Karargyris A, Syed AB, Boyko O, Moradi M, et al. Automatic bounding box annotation of chest X-ray data for localization of abnormalities. In: 2020 IEEE 17th international symposium on biomedical imaging. IEEE; 2020, p. 799–803.
[55] Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, et al. CheXpedition: investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. 2020, arXiv preprint arXiv:2002.11379.
[56] Yan C, Yao J, Li R, Xu Z, Huang J. Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. 2018, p. 103–10.
[57] Guendel S, Grbic S, Georgescu B, Liu S, Maier A, Comaniciu D. Learning to recognize abnormalities in chest X-rays with location-aware dense networks. In: Progress in pattern recognition, image analysis, computer vision, and applications: 23rd iberoamerican congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, proceedings 23. Springer; 2019, p. 757–65.
[58] Li Z, Wang C, Han M, Xue Y, Wei W, Li LJ, et al. Thoracic disease identification and localization with limited supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8290–9.
[59] Xu M, Zhang Z, Wei F, Hu H, Bai X. Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 2945–54.
[60] Cheng B, Misra I, Schwing AG, Kirillov A, Girdhar R. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 1290–9.
[61] Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
[62] Zhang K, Liu D. Customized segment anything model for medical image segmentation. 2023, arXiv preprint arXiv:2304.13785.
[63] Wu JT, Agu NN, Lourentzou I, Sharma A, Paguio JA, Yao JS, et al. Chest imagenome dataset for clinical reasoning. 2021, arXiv preprint arXiv:2108.00316.
[64] Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017, arXiv preprint arXiv:1711.05101.
[65] Xiao J, Bai Y, Yuille A, Zhou Z. Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023, p. 3588–600.
[66] Zhang Y, Luo L, Dou Q, Heng PA. Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. Med Image Anal 2023;86:102772.
[67] Kim E, Kim S, Seo M, Yoon S. XProtoNet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 15719–28.
[68] Rymarczyk D, Struski Ł, Górszczak M, Lewandowska K, Tabor J, Zieliński B. Interpretable image classification with differentiable prototypes assignment. In: European conference on computer vision. Springer; 2022, p. 351–68.
[69] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019, p. 6105–14.
[70] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021, p. 8748–63.