Interpreting Radiologist's Intention from Eye Movements in Chest X-ray Diagnosis

Trong Thang Pham
University of Arkansas
USA

Anh Nguyen
University of Liverpool
UK

Zhigang Deng
University of Houston
USA

Carol C. Wu
MD Anderson Cancer Center
USA

Hien Nguyen
University of Houston
USA

Ngan Le
University of Arkansas
USA

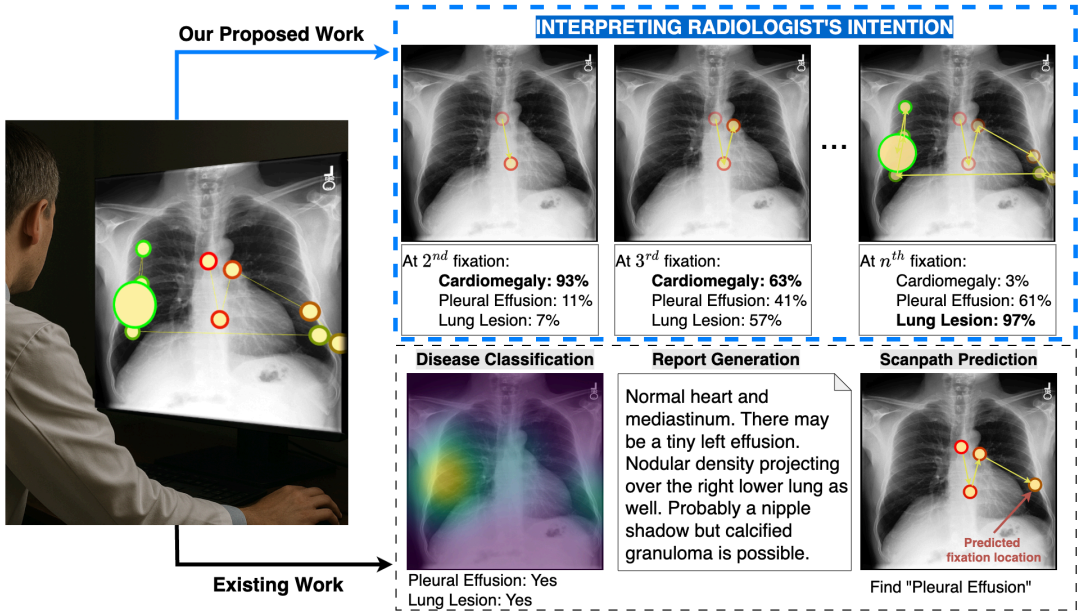


Figure 1: From existing gaze datasets (e.g., EGD and REFLACX) that are illustrated in the left image, existing research on gaze-assisted medical AI primarily focus on developing systems to assist radiologists by performing tasks such as disease classification, report generation, or mimicking visual search patterns through scanpath prediction tasks. Our work extends apart from these tasks and focuses on understanding radiologists by interpreting the radiologist's intention behind each of their captured gaze points.

Abstract

Radiologists rely on eye movements to navigate and interpret medical images. A trained radiologist possesses knowledge about the potential diseases that may be present in the images and, when searching, follows a mental checklist to locate them using their gaze. This is a key observation, yet existing models fail to capture the underlying intent behind each fixation. In this paper, we introduce a deep learning-based approach, *RadGazeIntent*, designed to model this behavior: having an intention to find something and actively searching for it. Our transformer-based architecture processes both the temporal and spatial dimensions of gaze data, transforming fine-grained fixation features into coarse, meaningful representations of diagnostic intent to interpret radiologists' goals. To capture the nuances of radiologists' varied intention-driven behaviors, we process existing medical eye-tracking datasets to create three intention-labeled subsets: RadSeq (Systematic Sequential Search), RadExplore (Uncertainty-driven Exploration), and RadHybrid (Hybrid Pattern). Experimental results demonstrate RadGazeIntent's ability to predict which findings radiologists are examining at specific moments, outperforming baseline methods across all intention-labeled datasets.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Health care information systems**.

Keywords

Eye Gaze Data, Deep Learning, Medical Image Analysis, Radiologist's Intention

ACM Reference Format:

Trong Thang Pham, Anh Nguyen, Zhigang Deng, Carol C. Wu, Hien Nguyen, and Ngan Le. 2018. Interpreting Radiologist's Intention from Eye Movements in Chest X-ray Diagnosis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Radiologists rely on eye movements to navigate and interpret medical images, leading to a natural question: "what are they trying to find at this moment?" Most current works attempt to mimic radiologists to create digital twins rather than using data from them to understand them, which could be crucial in interactive systems. To the best of our knowledge, no existing work deciphers the underlying intentions behind each fixation in medical imaging analysis.

As shown in Table 1, existing Aritificial Intelligence approaches (AI) fall short in intention prediction. I-AI [38], despite designing models to mimic radiologist decision-making processes, only focuses on predicting heatmaps (spatial information) without explaining the purpose of each fixation. Similarly, EyeXNet [20], while using fixation data, takes full heatmaps as input without considering temporal characteristics, and is used for localization rather than explaining fixation meaning. The work most similar to ours is ChestSearch [39], which predicts how a radiologist would view an chest X-ray (CXR) image given a specific finding to search for. However, it is limited to a single finding, whereas in reality, radiologists look for multiple findings, and each gaze point does not always serve only one class, leading to complex relationships. Furthermore, the GazeSearch data from ChestSearch significantly reduce fixation points from the original data, which would be difficult to implement in real scenarios for an interactive system. In contrast, our RadGazeIntent model directly takes the fixation sequence and chest X-ray image to predict which finding each fixation is used to identify, along with confidence scores.

Because we strive to understand radiologists' intentions, we designe our model based on the premise that eye movements [36] provide a window into cognitive processes. We conceptualize intention through three complementary perspectives as illustrated in Figure 1: First, intention may manifest as a systematic sequential search where radiologists follow a mental checklist, targeting specific findings sequentially and perform medical visual search tasks sequentially [39]. Second, intention could reflect uncertainty-driven exploration, where radiologists respond opportunistically to visual cues without predetermined targets, in practice, this means that for a predefined set of findings, radiologists identify and report whatever they happen to observe [12]. Third, intention might follow a hybrid pattern where radiologists initially conduct a brief reference scan of the entire image before focusing intently on a single pathology, effectively combining the systematic and opportunistic approaches [7]. In this paper, we cannot completely rule out any of these possibilities, so we process the original data and evaluate all methods across multiple definitions to provide the most comprehensive perspective. To support these different interpretations of radiologist behavior, we introduce three corresponding datasets: RadSeq (Systematic Sequential Search), RadExplore (Uncertainty-driven Exploration), and RadHybrid (Hybrid Pattern). Each dataset represents a different conceptualization of how radiologists allocate their visual attention during diagnosis.

To model the intention behind each fixation, we use a transformer-based architecture called RadGazeIntent to model gaze sequences with three major characteristics: incorporating both peripheral and foveal information to mimic the visual information humans perceive [51], ensuring fixation information adheres to causality (earlier fixations cannot access information from later ones), and recognizing that fixations are not independent but complementary to each other, meaning adjacent fixations can be combined into more complex features. This approach allows our model to learn patterns from "fine-grained and noisy" fixation data and transform them into "coarse and abstract representations that cluster only the most relevant information." Our RadGazeIntent model employs a transformer architecture with a pooling mechanism to achieve this transformation.

Our main contributions include:

- **Benchmark:** Three new benchmark datasets, RadSeq, RadExplore, and RadHybrid, representing different conceptualizations of radiologist's intention.
- **RadGazeIntent:** A novel framework, RadGazeIntent, for classifying radiologists' fixations according to their underlying intentions, bridging the gap between visual search patterns and diagnostic reasoning.
- **Evaluation:** A comprehensive evaluation across multiple settings for predicting radiologist intentions, demonstrating our model's ability to generalize across varied intention definitions and consistently outperform baseline approaches.

2 Related Work

2.1 Gaze-Assisted Medical AI

General gaze prediction models have been on a rise from both static saliency prediction [1, 6, 8, 15, 17, 22, 24, 27] and scanpath prediction [9, 10, 13, 26, 28, 35, 40, 43, 51–53]. For example, DeepGaze III [26] and Gazeformer [35] use deep learning to predict scanpaths in free-viewing tasks. Chen et al. [11] advance this with individualized scanpath prediction. However, these methods lack adaptation to medical contexts.

Recent approaches integrate medical gaze data into AI frameworks. Karargyris et al. [25] introduce a chest X-ray dataset with eye-tracking, but did not align gaze patterns with diagnostic labels. I-AI [38] is a system decoding radiologists' focus, primarily mimicking expert attention by predicting gaze heatmap. Pham et al. [39] then introduce GazeSearch for radiology findings search, focusing on scanpath prediction rather than intention.

A key limitation across these efforts is the absence of explicit alignment between gaze sequences and diagnostic intention. They do not interpret *why* a radiologist fixates on a region, whether for

Table 1: Comparison of previous works on eye tracking assistance methods. Most existing works using eye tracking datasets primarily utilize information in heatmap form (spatial modeling) to solve problems within their corresponding settings. Recently, ChestSearch [39] has focused on the visual search problem and proposed models to incorporate temporal information. It is also notable that a temporal classifier proposed by Karargyris et al. [25] can model temporal aspects by using RNNs on heatmaps of gaze sequences. However, despite having both temporal and spatial modeling capabilities, no existing work has tackled the problem of understanding the intention behind each gaze point.

Methods	Temporal Modeling	Spatial Modeling	Intention Interpretation	Tasks
I-AI [38]	✗	✓	✗	*Disease Classification*
GazeRadar [4]	✗	✓	✗	*Disease Localization*
EGGCA-Net [37]	✗	✓	✗	*Report Generation*
EyeXNet [20]	✗	✓	✗	*Disease Localization*
Karargyris et al. [25]	✓	✓	✗	*Disease Classification*
ChestSearch [39]	✓	✓	✗	*Scanpath Prediction*
RadGazeIntent (Ours)	✓	✓	✓	*Intention Interpretation*

systematic search, uncertainty, or hybrid strategies as we do in our framework. Moreover, Neves et al. [36] provide a review of gaze-driven interpretability in radiology, also confirming this gap and calling for AI models that decode expert intent, a challenge that our work directly addresses.

2.2 Multi-Label Findings Classification

Multi-label classification of radiological findings has seen significant progress with deep learning [2, 18, 21, 29, 31–33, 41, 44, 45, 50, 54], particularly for chest X-rays. CheXNet [41] demonstrated radiologist-level pneumonia detection using a convolutional neural network (CNN), while Irvin et al. [23] introduced CheXpert, a large dataset with uncertainty labels for multi-label classification. Wu et al. [48] further advanced this with the Chest ImaGenome dataset, incorporating clinical reasoning via anatomical annotations. Transformer-based models have also been applied to this task. Taslimi et al. [44] used Swin Transformers for multi-label chest X-ray classification, achieving robust performance across findings. Wang et al. [47] proposed a multi-granularity cross-modal alignment framework, integrating text and image features for generalized representation learning. These works excel in identifying multiple pathologies but rely solely on image data, ignoring gaze-informed supervision that could reveal diagnostic priorities.

*Unlike these approaches, our model integrates gaze sequences with multi-label classification, focusing on predicting not **what** findings are present but **which** finding a radiologist is examining at a given moment.*

3 Methodology

3.1 Problem Formulation

We formulate the task of fixation-based intention interpretation as a sequence labeling problem. Given a series of T eye fixations $\mathcal{F} = \{f_1, f_2, \ldots, f_T\}$ where each fixation $f_i = (x_i, y_i, d_i)$ consists of spatial coordinates (x_i, y_i) and duration d_i. Each fixation may be associated with one of K possible intentions, reflecting the observation that multiple consecutive fixations typically correspond to a single cognitive process. We denote the ground truth intention label for each fixation as $L = \{l_1, l_2, \ldots, l_T\}$ where $l_i \in \{0, 1\}^K$. Our goal is to identify the underlying user intentions that generated these fixations.

3.2 Architecture: RadGazeIntent

The proposed framework processes a CXR through a sequential pipeline as shown in Figure 2: Starting with CXR images and fixation data, the Feature Extraction module generates peripheral and foveal feature maps. These features are passed to the Structural Embedding module, which encodes spatial and temporal characteristics of the fixation sequence. The embedded features are then processed by the Peripheral-aware Causal Self-Attention mechanism to integrate contextual dependencies. Finally, the Intention Decoder classifies the diagnostic purpose of each fixation and outputs corresponding confidence scores for relevant radiological findings.

Feature Extraction. The goal of this step is to create a feature pyramid to represent both peripheral visual information and fovea visual information of an image. For an input image $I \in \mathbb{R}^{H \times W}$, we use a Feature Pyramid Network (FPN) [30] with ResNet [19] backbone to obtain pyramid features $\{P^1, P^2, P^3, P^4\}$ with varying resolutions. We select $P^l = P^1 \in \mathbb{R}^{C \times H/32 \times W/32}$ with the lowest resolution to represent peripheral visual information and $P^h = P^4 \in \mathbb{R}^{C \times H/4 \times W/4}$ with the highest resolution to represent fovea visual information.

Structural Embedding. This component embeds the fixation sequence into meaningful features that represent: (1) feature relevance to image content, (2) 2D spatial properties, (3) temporal sequence information. First, we perform Fovea Mapping by extracting features from the corresponding spatial locations (x_i, y_i) in the fovea feature map P^h, mapping a point (x, y) to the feature at block $(x/4, y/4)$ based on the scale of P^h. In total, we obtain $E_f = \{e_1, e_2, \ldots, e_T\} \in \mathbb{R}^{C \times T}$ corresponding to T input fixations, wehre $e_i \in \mathbb{R}^C$ is the embedded feature of the i^{th} fixation.

Then, we apply 2D Spatial Embedding [14] to embed 2D spatial information into the feature representation using sinusoidal functions, followed by 1D Temporal Embedding [46] which incorporates sequence order information by encoding the sequential position t of each fixation point. After embedding, we obtain structural embedded fixation feature $\bar{E}_f$. Simultaneously, we flatten P^l

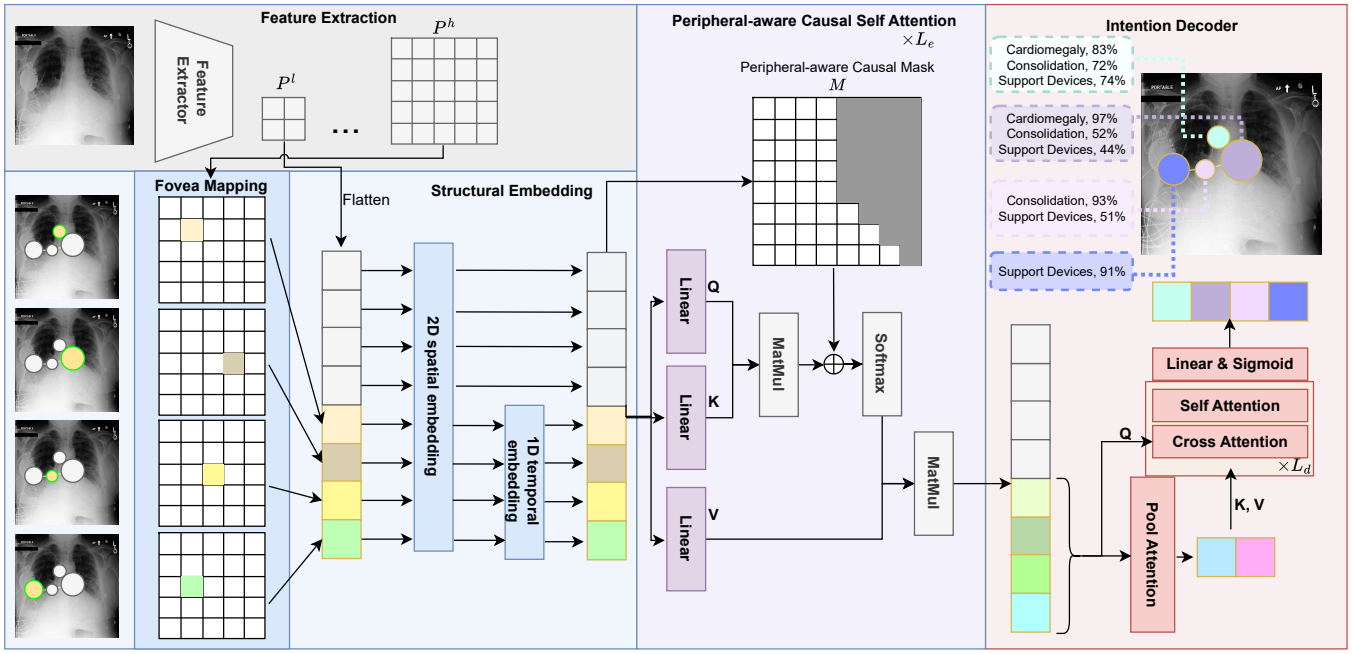


Figure 2: Overall framework of RadGazeIntent. RadGazeIntent aims to analyze medical image fixation patterns to determine diagnostic intentions. The Feature Extraction module processes a input CXR to create two distinct feature maps, peripheral (P^l) for general context and fovea (P^h) for detailed focus areas. The Structural Embedding module transforms fixation coordinates into feature representations using the fovea feature map (P^h). This module incorporates both spatial (2D) and temporal (1D) structural information to maintain the relationship between fixation points. Then the Peripheral-aware Causal Self Attention is a specialized attention mechanism that enables the model to learn features for each fixation in context. It references peripheral-level image information while preserving causality, i.e. ensuring that earlier fixations cannot access information from later ones, with the help of Peripheral-aware Causal Mask M (see Eq. (1)). Our implementation stacks L_e blocks of this attention mechanism. The Intention Decoder uses a Pool Attention module to condense the token representation. These condensed tokens are processed through L_d blocks of cross-attention and self-attention layers, followed by linear and sigmoid layers. This produces confidence scores for specific diagnostic findings, such as "Cardiomegaly?, "Consolidation" and "Support Devices," as shown in the top right of this figure.

into multiple tokens $\bar{P}^l \in \mathbb{R}^{C\times(HW/1024)}$ and also apply 2D spatial embedding to these tokens. We then concatenate the embeddings into a single vector $E_s = [\bar{P}^l, \bar{E}_f] \in \mathbb{R}^{C\times(HW/1024+T)}$ and pass it to the next step.

Peripheral-aware Causal Self Attention. This specialized self-attention mechanism incorporates peripheral information while maintaining causality in the sequence processing. It uses a Peripheral-aware Causal Mask that allows each fixation to access all preceding fixations and peripheral information but prevents access to future fixations. As we have a total of $HW/1024 + T$ tokens, we need a mask M with the size of $(HW/1024 + T) \times (HW/1024 + T)$. We create Peripheral-aware Causal Mask as:

$$\mathrm{M}_{ij} = \begin{cases} 0 & \text{if } i > j \text{ or } j \in [1, HW/1024] \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

where $i, j \in [1, HW/1024 + T]$ corresponding to row and column indexes.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathrm{M}}{\sqrt{d_k}}\right)V \tag{2}$$

where $\mathbf{Q}$ is the query, $\mathbf{K}$ is the key, $\mathbf{V}$ is the value, created by passing E_s through separate Linear layers [46], M is the Peripheral-aware Causal Mask, and d_k is the hidden dimension of $\mathbf{K}$.

This self-attention block transforms our encoded features into a deep latent space representing both image and fixation information simultaneously. Multiple (L_e) layers of this attention mechanism are applied to capture complex relationships. After this step, we obtain the contextualized feature $E'_s = [\bar{P}^{l'}, E'_f]$, where $\bar{P}^{l'}$ is a global information and E'_f is the fixation feature. In this module, $\bar{P}^{l'}$ aims to enrich context information to E'_f, so we split and only use E'_f for the next step.

Intention Decoder. The role of the Intention Decoder is to explicitly filter out noise and capture more complex patterns that represent intentions and then decode that features into the intention behind each fixation. Intuitively, intentions typically span multiple fixations. Thus, we use Pool Attention to compress the feature sequence E'_f, reducing from T tokens to fewer tokens, E^*_f . Next, Self-Attention and Cross-Attention layers allow each feature token

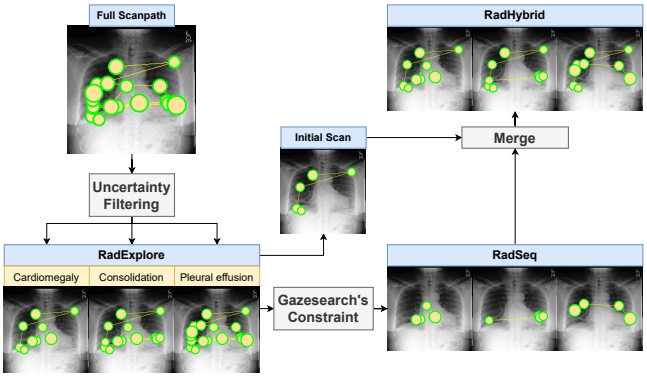
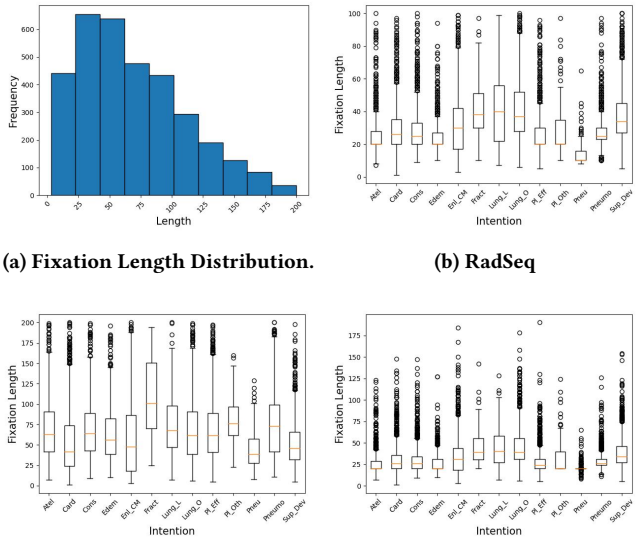


Figure 3: Illustration of our dataset creation process, which transforms the original eye-tracking data into three distinct experimental settings: RadSeq (Systematic Sequential Search), RadExplore (Uncertainty-driven Exploration), and RadHybrid (Hybrid Pattern). Beginning with a full scanpath showing radiologist fixation points (green and yellow) on a chest X-ray, we apply *uncertainty filtering* to isolate fixations that fall outside annotated findings, forming RadExplore, which models exploratory behavior under diagnostic uncertainty. In this example, we extract fixations for three findings: Cardiomegaly, Consolidation, and Pleural Effusion. Next, we implement Gazesearch's constraints [39] to convert RadExplore into pathology-focused fixations, creating RadSeq, which simulates systematic and targeted search. To construct RadHybrid, we merge the extracted *initial scan* from the first few seconds (see Section 4.1) with RadSeq to capture a behavior pattern that begins with broad scanning and transitions into focused searching.



(a) Fixation Length Distribution. **(b) RadSeq**

(c) RadExplore **(d) RadHybrid**

Figure 4: Statistical analysis of eye fixation patterns across different datasets. (a) Histogram showing the overall distribution of fixation lengths, with most fixations concentrated between 25-75 points. (b-d) Box plots displaying fixation length distributions across 13 radiological findings for three different datasets: RadSeq, RadExplore, and RadHybrid. Each box plot represents the median, interquartile range, and outliers of fixation lengths for specific medical findings. The abbreviated labels correspond to: Atel (Atelectasis), Card (Cardiomegaly), Cons (Consolidation), Edem (Edema), Enl_CM (Enlarged Cardiomediastinum), Fract (Fracture), Lung_L (Lung Lesion), Lung_O (Lung Opacity), Pl_Eff (Pleural Effusion), Pl_Oth (Pleural Other), Pneu (Pneumonia), Pneumo (Pneumothorax), and Sup_Dev (Support Devices).

from E'_f to query and select the most appropriate latent features representing underlying intentions from E^*_f. This process is repeated for L_d layers. Finally, a Linear layer transforms the decoder output into the intention space, and a Sigmoid layer normalizes values between 0 and 1, providing confidence scores for specific findings such as "Cardiomegaly," "Consolidation," and "Support Devices" as shown in the Figure 2.

Objective Function. The proposed problem of fixation interpretation can be formulated as a multi-label classification task, so we use binary cross-entropy as our loss function:

$$\mathcal{L} = -\frac{1}{TK}\sum_{i=1}^{T}\sum_{k=1}^{K}\left[l_{ik}\log(\hat{l}_{ik}) + (1-l_{ik})\log(1-\hat{l}_{ik})\right] \quad (3)$$

where l_i is the ground truth label, $\hat{l}_i$ is the predicted probability.

4 Experiments

4.1 Datasets

To investigate the intention behind each radiologist's gaze, we derive three intention-labeled datasets by post-processing two publicly available gaze datasets, EGD [25] and REFLACX [5], under distinct behavioral assumptions that reflect plausible visual search strategies [3, 12, 39]. Figure 3 illustrates the overview of our data processing steps. The three newly introduced intention-labeled datasets are as follows:

RadExplore: Uncertainty-Driven Exploration. This dataset considers intention as opportunistic visual search [12], assuming radiologists do not follow a fixed order and may consider all report findings simultaneously. This reflects maximal ambiguity: all fixations are potentially relevant to any finding, leaving intention disambiguation to later modeling stages. Formally, let $S = \{s_1, s_2, ..., s_{|S|}\}$ be the sequence of sentences in the transcript, where each sentence has an end time s^e_j, and $\{\tau_i\}_{i=1}^{T}$ is the set of captured timestampt for $\{f_i\}_{i=1}^{T}$ fixations. We use CheXbert [42] to find the corresponding class for all sentences and produce $C = \{c_1, c_2, \ldots, c_{|S|}\}$. Then we compute the ground truth labels as:

$$\mathbf{l}_{ik} = \begin{cases} 1 & \text{if there exists } j \in [1, |S|] \text{ such that } \tau_i \le s^e_j \text{ and } k = c_j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $i \in [1, T]$ indexes fixations and $k \in [1, K]$ indexes intention labels. We refer to this step as Uncertainty Filtering.

RadSeq: Systematic Sequential Search. This dataset assumes radiologists follow a sequential checklist of findings [39] and solely focus on searching clues for a particular finding at a time. Gazesearch's constraints [39] comprise of two procedures: radius-based filtering and time-spent constraining. Using these constraints on RadExplore, we obtain the beginning time beg_k and end time end_k for all K intentions in the report S. Then we compute the ground truth labels as:

$$\mathbf{l}_{ik} = \begin{cases} 1 & \text{if } \tau_i \in [\text{beg}_k, \text{end}_k] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Unlike the original paper [39], we set the radius to zero in the radius-based filtering procedure to avoid discarding any fixation points, thereby preventing a reduction in the temporal and spatial information of the fixations.

RadHybrid: Hybrid Pattern. According to [3], a radiologist's intention leads to a two-phase search process where radiologists begin with a broad overview and later narrow focus to specific pathologies. To get the "broad overview" behavior, we extract the scanning fixations within the first τ^* seconds. Often, τ^* is 1 second according to [7]. Finally, we merge initial scanning fixations with RadSeq:

$$l_{ik} = \begin{cases} 1 & \text{if } \tau_i \leq \tau^* \text{ or } \tau_i \in [\text{beg}_k, \text{end}_k] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\tau^* = 1$ is the initial scanning time.

Each of the three datasets, RadSeq, RadExplore, and RadHybrid, is derived from the same source of eye tracking data, i.e., 1,079 samples from EGD and 2,483 samples from REFLACX. Figure 4 illustrates the patterns in radiologists' eye fixation behavior across different radiological findings. The histogram shows most fixations are relatively brief (25-75 points), though with some longer outliers. Across all three datasets (RadSeq, RadExplore, and RadHybrid), certain conditions consistently demand longer fixation lengths, particularly Lung Lesions and Fractures, which stand out with higher median values and wider interquartile ranges, suggesting these conditions may require longer visual attention during diagnosis.

4.2 Compared Baselines

We compare our approach against several baselines representing different architectural paradigms: We compare our model against a range of existing approaches that capture different modeling assumptions for intention prediction from radiologist gaze data.

MLP. This baseline uses a multilayer perceptron that processes each fixation independently. Each input fixation is first mapped with fovea feature like in our framework. Then they are passed through three fully connected layers (512, 256, K units) with ReLU activations, where K denotes the number of intention classes. Finally the latent features are passed through a sigmoid layer for multi-label classification. This model lacks any temporal modeling or spatial aggregation over time.

LSTM. A sequential model that encodes temporal dynamics in gaze behavior using a unidirectional LSTM. Similar to MLP, we also use mapped fovea features to represent the fixation token features. Then we use the LSTM decoder on the fixation token features. The model has 256 hidden units with a dropout rate of 0.2 between layers. The final hidden state is projected to the intention space via a fully connected layer. Finally the projected features are passed through a sigmoid layer for multi-label classification.

Karargyris et al. [25]. This model first transforms fixation sequences into spatial heatmaps using Gaussian kernels. The input CXR and these heatmaps are passed through a ResNet-18 CNN encoder, followed by a bidirectional LSTM (256 hidden units), a temporal convolutional layer, and a final classification head. Unlike the original implementation in [25], we modify the final classification head: instead of predicting three classes, we apply the classification head separately for each token to get the multi-label prediction.

ChestSearch [39]. This baseline is originally designed for the GazeSearch dataset. We change the final decoder heads of ChestSearch from decoding heatmaps to a classifier for predicting intention.

All deep learning models are trained with the Adam optimizer, initial learning rate of 1e-4 with cosine annealing schedule, and for the same number of epochs (100) with early stopping based on validation performance to ensure fair comparison.

4.3 Implementation Details

We use a Pyramid Feature Network with ResNet-50 backbone [19] as the Feature Extractor, initialized from a checkpoint pre-trained using MGCA [47] for 50 epochs with a batch size of 144. This Feature Extractor is frozen when we train the full pipeline. We stack $L_e = 4$ Peripheral-aware Causal Self Attention layers with hidden size $D = 384$ and $H = 8$ attention heads.

To reduce sequence length and retain complex features, we apply a pooling attention layer [16] with stride of 2 and kernel size of 5 tokens. The Intention Decoder contains $L_d = 6$ blocks of self-attention and cross-attention. We train the entire model for 4,000 iterations using the AdamW optimizer [34], with a learning rate of 1×10^{-5} and a batch size of 32. All experiments are conducted on a single NVIDIA A6000 GPU with 48GB of RAM.

We then evaluate intention predictors using a set of classification metrics, i.e., Accuracy (ACC), F1-score (F1), Precision (P), and Recall (R), for every pair of fixation-intention. We run 5-fold cross validation and report 95% confidence interval in Section 4.4.

4.4 Quantitative Results

Table 2 presents the quantitative performance of our proposed model compared to four baseline methods (MLP, LSTM, Karargyris et al., and ChestSearch) across two eye-tracking sources (EGD and REFLACX), evaluated under three datasets representing different intention perspectives: RadExplore, RadSeq, and RadHybrid.

In RadSeq, the higher F1-scores (72.05% for EGD and 69.87% for REFLACX) reflect the model's ability to accurately capture the radiologist's systematic sequential scanning. Baseline models, particularly simpler ones like MLP and LSTM, struggle to model the temporal order of fixations, resulting in lower accuracy and recall. Thanks to our transformer-based architecture, RadGazeIntent effectively models these sequential dependencies.

As shown in Figure 4, RadExplore exhibits much longer fixation sequences than RadSeq, requiring a greater number of predictions. Despite this, *RadGazeIntent* continues to demonstrate robustness. For instance, its precision scores (72.25% for EGD and 70.89% for

Table 2: Performance comparison across two datasets (EGD and REFLACX) with 95% confidence intervals (±). Metrics include Accuracy (ACC), F1-score (F1), Precision (P), and Recall (R).

Datasets \ Data Sources	Model	EGD				REFLACX			
		ACC (%)	F1 (%)	P (%)	R (%)	ACC (%)	F1 (%)	P (%)	R (%)
RadSeq	MLP	73.87 (±1.3)	49.14 (±1.6)	62.16 (±1.5)	51.40 (±1.8)	82.55 (±1.2)	52.33 (±1.4)	58.90 (±1.6)	54.76 (±1.3)
	LSTM	81.23 (±1.5)	56.77 (±1.4)	59.12 (±1.3)	54.89 (±1.7)	79.98 (±1.7)	55.21 (±1.5)	60.01 (±1.3)	53.43 (±1.4)
	Karargyris et al.	84.02 (±1.2)	61.88 (±1.2)	64.77 (±1.1)	59.45 (±1.0)	81.22 (±1.1)	59.34 (±1.2)	63.15 (±1.1)	56.70 (±1.3)
	ChestSearch	87.35 (±1.0)	68.20 (±1.0)	70.11 (±1.2)	66.30 (±1.1)	85.02 (±1.0)	65.44 (±1.0)	68.21 (±1.2)	63.70 (±1.1)
	Ours	**88.85 (±0.7)**	**72.05 (±0.9)**	**74.01 (±1.0)**	**70.51 (±0.9)**	**86.92 (±0.9)**	**69.87 (±0.9)**	**72.12 (±1.0)**	**67.90 (±0.9)**
RadExplore	MLP	72.45 (±1.4)	50.32 (±1.5)	59.20 (±1.6)	52.10 (±1.3)	81.00 (±1.3)	51.44 (±1.5)	56.70 (±1.3)	50.33 (±1.4)
	LSTM	80.33 (±1.6)	53.88 (±1.3)	57.78 (±1.5)	52.04 (±1.4)	78.45 (±1.5)	54.11 (±1.3)	58.20 (±1.3)	51.80 (±1.3)
	Karargyris et al.	83.12 (±1.2)	60.45 (±1.1)	62.99 (±1.1)	58.77 (±1.0)	80.76 (±1.3)	57.44 (±1.3)	61.02 (±1.2)	55.21 (±1.1)
	ChestSearch	86.44 (±0.9)	66.10 (±1.1)	68.30 (±1.2)	64.55 (±1.0)	84.01 (±1.0)	63.11 (±1.0)	66.90 (±1.1)	61.22 (±1.0)
	Ours	**87.95 (±0.8)**	**70.14 (±0.9)**	**72.25 (±1.0)**	**68.01 (±0.9)**	**85.40 (±0.8)**	**67.33 (±0.9)**	**70.89 (±1.0)**	**65.92 (±0.9)**
RadHybrid	MLP	72.12 (±1.3)	48.99 (±1.4)	60.45 (±1.5)	50.77 (±1.3)	80.01 (±1.3)	50.22 (±1.4)	55.10 (±1.2)	49.85 (±1.4)
	LSTM	79.80 (±1.5)	52.77 (±1.3)	55.89 (±1.4)	50.43 (±1.5)	77.33 (±1.4)	51.10 (±1.3)	56.01 (±1.2)	48.88 (±1.2)
	Karargyris et al.	83.56 (±1.1)	58.60 (±1.1)	61.90 (±1.2)	56.45 (±1.0)	79.66 (±1.2)	56.77 (±1.1)	60.44 (±1.3)	54.90 (±1.1)
	ChestSearch	86.22 (±0.9)	65.77 (±1.0)	67.91 (±1.1)	63.80 (±1.0)	83.45 (±0.9)	62.70 (±1.0)	65.11 (±1.0)	60.88 (±1.0)
	Ours	**88.21 (±0.8)**	**71.11 (±0.8)**	**73.20 (±0.9)**	**69.88 (±0.8)**	**86.02 (±0.8)**	**68.44 (±0.9)**	**71.55 (±0.9)**	**66.78 (±0.8)**

REFLACX) suggest that our model better infers the intent behind each fixation, even amidst uncertainty, compared to the baselines.

Finally, in RadHybrid, our model consistently outperforms all baselines across both datasets, with notable gains in recall (69.88% for EGD and 66.78% for REFLACX), highlighting its ability to capture both broad overview and focus scanning phases. Baseline models, except ChestSearch, lack mechanisms to disentangle these phases, leading to lower F1-scores as they conflate coarse and fine-grained fixations. Our model's improvement over ChestSearch is attributed to its pooling mechanism, which enables the separation of exploratory and focused patterns from the input fixations.

Overall, the quantitative results validate the effectiveness of our framework across all three intention definitions, establishing a strong benchmark for intention interpretation.

4.5 Qualitative Results

In this section, we present a qualitative analysis of our model's ability to interpret radiologists' fixation patterns and predict their underlying intentions during medical image analysis.

Figure 5 presents a qualitative comparison of radiologist intention prediction results across three experimental settings and various intentions. Our proposed model demonstrates superior performance compared to both baseline approaches (Karargyris et al. and ChestSearch), with prediction closely resembling ground truth. This is evidenced by the prevalence of green fixation points (correct predictions) in our model's results across all settings, while the competing approaches show considerably more red indicators (incorrect predictions). RadExplore and RadHybrid display denser fixation sequences overall, but our performance is still maintained and remains similar to the case with the fewest points, which is RadSeq. These qualitative results reinforce the quantitative advantage of our approach in accurately predicting radiologists' diagnostic intentions.

4.6 Ablation Study

Table 3 presents an ablation study to assess the impact of removing key components from our full model, providing insights into the importance of each element in capturing radiologists' gaze intentions. The study examines the effects of removing Pool Attention, 1D Temporal Embedding, 2D Spatial Embedding, Peripheral Features, and Fovea Mapping, highlighting their individual roles in the model's ability to accurately interpret gaze patterns.

w/o Pool Attention. In this setting, we remove the Pool Attention layer and use E'_f directly in the Cross Attention module. This leads to a moderate drop in performance (1.29–2.81%), highlighting its role in aggregating features that capture overall intent patterns across fixations.

w/o 1D Temporal Embedding. We remove the 1D Temporal Embedding block from the Structural Embedding module. Eliminating temporal encoding removes explicit modeling of fixation order, thereby discarding the temporal dynamics of visual attention. This results in a significant performance decline (4.34–4.68%), confirming that temporal progression is vital for understanding intention.

w/o 2D Spatial Embedding. We remove the 2D Spatial Embedding block from the Structural Embedding module. Without spatial embedding, the model no longer receives location-aware cues about where fixations occur on the image. The performance drop (3.24–6.67%) indicates that spatial context is also important.

w/o Peripheral Feature. This configuration excludes the coarse-resolution features P^l. As a result, the Peripheral-aware Causal Self Attention blocks perform causal self-attention solely on fixation features, without incorporating the broader contextual features from P^l. The performance drops (4.76–6.67%) suggest that peripheral information, though less detailed than foveal features, contributes to understanding scene context and supports intention interpretation.

w/o Fovea Mapping. In this setting, we replace the Fovea Mapping with a 2D layout embedding based on coordinates [49], commonly used in document understanding. This effectively removes the high-resolution foveal features P^h and replaces them with standard 2D spatial embeddings. This leads to the most severe degradation (11.73–12.12%), indicating that high-resolution features are crucial for identifying subtle diagnostic cues, especially during fine-grained examination phases. Their absence impairs the model's ability to localize fixation intent around medically salient regions.

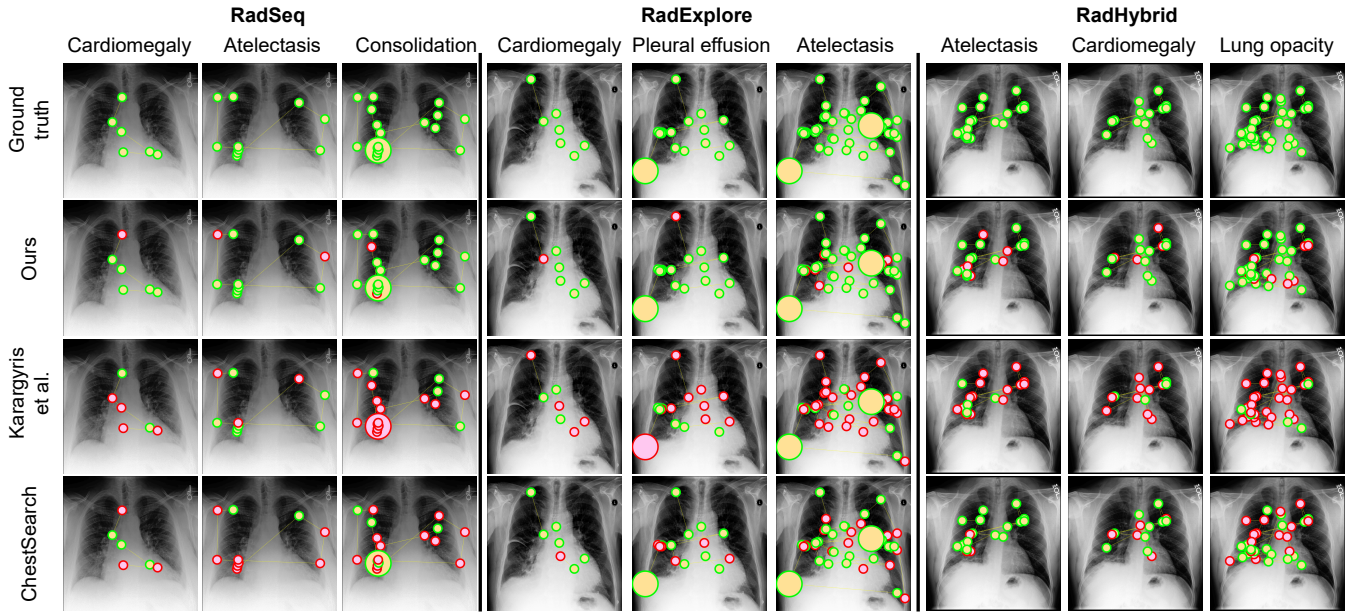


Figure 5: Qualitative comparison of radiologist intention prediction results across three datasets and various intention classes. The visualization presents chest X-rays with overlaid fixation points organized in a matrix format where columns represent different pathological findings across datasets, and rows represent different prediction methods: Ground truth, Our proposed model, Karargyris et al.'s approach, and ChestSearch. The input for the model is the full fixation sequence, and model's objecive is to predict the intention class for each point. ◯ represents correct prediction (i.e., confidence score of a fixation is greater than 0.5) and ◯ denotes incorrect prediction.

These results demonstrate that spatial and temporal cues, along with multi-scale visual processing (foveal and peripheral), are essential for effectively interpreting radiologists' gaze intentions.

Table 3: Ablation study on architectural components across different intention datasets. Metric: The average of F1-score (%) on both eye tracking sources (EGD and REFLACX).

Model Configuration	RadSeq	RadExplore	RadHybrid
Full Model	71.01	68.74	69.78
w/o Pool Attention	68.20	67.45	68.01
w/o 1D Temporal Embedding	66.33	64.10	65.44
w/o 2D Spatial Embedding	67.77	62.89	63.11
w/o Peripheral Feature	64.34	63.45	65.02
w/o Fovea Mapping	58.92	57.01	57.66

5 Conclusion

Discussion. In this work, we introduce a novel paradigm for interpreting radiologists' eye movements through the lens of intention, shifting the focus from mimicry-based modeling to a cognitively grounded understanding of visual search behavior. Our RadGazeIntent framework classifies each fixation point by its likely diagnostic purpose. We further propose three datasets, RadSeq, RadExplore, and RadHybrid, each representing a distinct observation of how radiologists allocate visual attention in practice. These settings allowed us to empirically test competing theories of gaze behavior in radiology, from structured checklist-based scanning to reactive exploration. Experiments show RadGazeIntent consistently outperforms existing baselines across all datasets.

Limitations. Despite these advances, several limitations remain. First, "intention" is inherently abstract and cannot be directly observed in eye tracking datasets, we infer intention from our observations on radiologist's behavior, which may not perfectly reflect cognitive processes. Second, while we model intention at the individual fixation level, not all fixations may map cleanly to a specific diagnostic objective (early or exploratory phases of scanning). This ambiguity introduces inherent noise that may confound classification. Finally, while conceptually diverse, our datasets focus exclusively on chest X-rays;; radiologists' behaviors in other imaging modalities, e.g., CT or MRI, may follow different patterns. Future work presents exciting opportunities to extend our work to other modalities, potentially revealing universal patterns of expert visual reasoning.

Broader Impact. By decoding the intent behind radiologists' gaze behavior, our approach opens new pathways for developing interactive, intention-aware systems that can collaborate with rather than replace human experts. Potential applications include gaze-guided report generation, intention-aware assistance during training, and real-time feedback systems that adapt to a user's diagnostic focus. Furthermore, understanding gaze intent has implications beyond radiology. For example, in surgical navigation, pathology slide review, and even education platforms that teach visual diagnostic skills.

References

[1] Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, and Sabine Susstrunk. 2023. TempSAL - Uncovering Temporal Information for Deep Saliency Prediction.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. 2021. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3478–3488.

[3] Raymond Bertram, Laura Helle, Johanna K Kaakinen, and Erkki Svedström. 2013. The effect of expertise on eye movement behaviour in medical image perception. *PloS one* 8, 6 (2013), e66169.

[4] Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. 2022. Gazeradar: A gaze and radiomics-guided disease localization framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 686–696.

[5] Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. 2022. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data* 9, 1 (2022), 350.

[6] Souradeep Chakraborty and otherss. 2022. Predicting Visual Attention in Graphic Design Documents. *IEEE Transactions on Multimedia (TMM)* (2022).

[7] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020. AiR: Attention with Reasoning Capability.. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[8] Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, and Junfeng He. 2023. Learning from Unique Perspectives: User-aware Saliency Modeling.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Predicting Human Scanpaths in Visual Question Answering.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] Xianyu Chen, Ming Jiang, and Qi Zhao. 2024. Beyond Average: Individualized Visual Scanpath Prediction.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[11] Xianyu Chen, Ming Jiang, and Qi Zhao. 2024. Beyond Average: Individualized Visual Scanpath Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25420–25431.

[12] Yupei Chen et al. 2022. Characterizing Target-Absent Human Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 5031–5040.

[13] Zhenzhong Chen and Wanjie Sun. 2018. Scanpath Prediction for Visual Attention using IOR-ROI LSTM.. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299.

[15] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing (IEEE TIP)* (2018).

[16] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6824–6835.

[17] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. 2020. Predicting Visual Importance Across Graphic Design Types.. In *ACM Symposium on User Interface Software and Technology*.

[18] Matej Gazda, Ján Plavka, Jakub Gazda, and Peter Drotar. 2021. Self-supervised deep convolutional neural network for chest x-ray classification. *IEEE Access* 9 (2021), 151972–151982.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. arXiv e-prints. *arXiv preprint arXiv:1512.03385* 10 (2015).

[20] Chihcheng Hsieh, André Luís, José Neves, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Joaquim Jorge, and Catarina Moreira. 2024. EyeXNet: Enhancing Abnormality Detection and Diagnosis via Eye-Tracking and X-ray Fusion. *Machine Learning and Knowledge Extraction* 6, 2 (2024), 1055–1071.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[22] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[23] Jeremy Irvin et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[24] Sen Jia and Neil D. B. Bruce. 2020. EML-NET:An Expandable Multi-Layer NETwork for Saliency Prediction. *Image and Vision Computing* (2020).

[25] Alexandros Karargyris et al. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data* 8, 1 (2021), 1–18.

[26] Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. 2022. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision (JoV)* (2022).

[27] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* (2016).

[28] Peizhao Li, Junfeng He, Gang Li, Rachit Bhargava, Shaolei Shen, Nachiappan Valliappan, Youwei Liang, Hongxiang Gu, Venky Ramachandran, Golnaz Farhadi, Yang Li, Kai J Kohlhoff, and Vidhya Navalpakkam. 2023. UniAR: Unifying Human Attention and Response Prediction on Visual Content. *arXiv preprint arXiv:2312.10175* (2023).

[29] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. 2018. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8290–8299.

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[31] Fengbei Liu, Yu Tian, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. 2021. 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, 426–436.

[32] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. 2019. 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*. 10632–10641.

[33] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging* 39, 11 (2020), 3429–3440.

[34] I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[35] Sounak Mondal et al. 2023. Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] José Neves, Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Anderson Maciel, Andrew Duchowski, Joaquim Jorge, and Catarina Moreira. 2024. Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. *European Journal of Radiology* (2024), 111341.

[37] Peixi Peng, Wanshu Fan, Yue Shen, Wenfei Liu, Xin Yang, Qiang Zhang, Xiaopeng Wei, and Dongsheng Zhou. 2024. Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics* (2024).

[38] Trong Thang Pham, Jacob Brecheisen, Anh Nguyen, Hien Nguyen, and Ngan Le. 2024. 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*. 7850–7859.

[39] Trong Thang Pham, Tien-Phat Nguyen, Yuki Ikebe, Akash Awasthi, Zhigang Deng, Carol C Wu, Hien Nguyen, and Ngan Le. 2024. GazeSearch: Radiology Findings Search Benchmark. *arXiv preprint arXiv:2411.05780* (2024).

[40] Mengyu Qiu, Yi Guo, Mingguang Zhang, Jingwei Zhang, Tian Lan, and Zhilin Liu. 2023. Simulating Human Visual System Based on Vision Transformer.. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*.

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

[42] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv:2004.09167 [cs.CL]

[43] Wanjie Sun, Zhenzhong Chen, and Feng Wu. 2019. Visual Scanpath Prediction using IOR-ROI Recurrent Mixture Density Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)* (2019).

[44] Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. 2022. SwinCheX: Multi-label classification on chest X-ray images with transformers. *arXiv preprint arXiv:2206.04246* (2022).

[45] Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, and Marcel Worring. 2023. Probabilistic Integration of Object Level Annotations in Chest X-ray Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3630–3640.

[46] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[47] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems* 35 (2022), 33536–33549.

[48] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316* (2021).

[49] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1192–1200.

[50] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. 2018. 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*. 103–110.

[51] Zhibo Yang et al. 2020. Predicting Goal-directed Human Attention Using Inverse Reinforcement Learning.. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[52] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. 2022. Target-absent Human Attention.. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[53] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. 2023. Predicting Human Attention using Computational Attention. *arXiv preprint arXiv:2303.09383v2* (2023).

[54] Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703* (2018).