GazeSearch: Radiology Findings Search Benchmark

Trong Thang Pham*, Tien-Phat Nguyen†, Yuki Ikebe*, Akash Awasthi‡, Zhigang Deng‡,
Carol C. Wu§, Hien Nguyen‡, and Ngan Le*
*University of Arkansas, Fayetteville, AR, USA
†University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
‡University of Houston, Houston, TX, USA
§MD Anderson Cancer Center, Houston, TX, USA

Abstract

Medical eye-tracking data is an important information source for understanding how radiologists visually interpret medical images. This information not only improves the accuracy of deep learning models for X-ray analysis but also their interpretability, enhancing transparency in decision-making. However, the current eye-tracking data is dispersed, unprocessed, and ambiguous, making it difficult to derive meaningful insights. Therefore, there is a need to create a new dataset with more focus and purposeful eye-tracking data, improving its utility for diagnostic applications. In this work, we propose a refinement method inspired by the target-present visual search challenge: there is a specific finding and fixations are guided to locate it. After refining the existing eye-tracking datasets, we transform them into a curated visual search dataset, called GazeSearch, specifically for radiology findings, where each fixation sequence is purposefully aligned to the task of locating a particular finding. Subsequently, we introduce a scan path prediction baseline, called ChestSearch, specifically tailored to GazeSearch. Finally, we employ the newly introduced GazeSearch as a benchmark to evaluate the performance of current state-of-the-art methods, offering a comprehensive assessment for visual search in the medical imaging domain. Code is available at https://github.com/UARK-AICV/GazeSearch.

1. Introduction

Artificial Intelligence (AI) has been growing rapidly and become an important part of daily life [3, 32, 34–36, 42, 46–49, 56, 74, 79], including important workers like clinical experts and healthcare providers [4, 25, 31, 37, 54, 65, 72]. Beyond achieving high performance, it is essential to develop AI systems that offer explainable and interpretable decision-making [2, 21, 23, 51, 52, 61, 63, 64, 73]. This is especially important in sensitive domains such as healthcare, where credibility and reliability are critical to ensuring trust and safe implementation. Even though human experts remain the ultimate authority in decision-making, researchers are focusing on improving AI-assisted systems to reduce the burden for the experts. For example, we can use AI to produce preliminary results and the experts can either confirm or adjust [72]. As a result, the collaborative approach between AI and professionals has successfully improved radiological diagnosis in many cases compared to radiologists or the system alone [65]. However, a key challenge is building trust in AI, especially with black-box models in healthcare, such as CXR analysis. This has increased the demand for models that mimic radiologists' behavior to improve interpretability. For instance, aligning AI systems with radiologists' visual attention patterns is essential [51, 55]. This has opened a new domain of research focused on model-

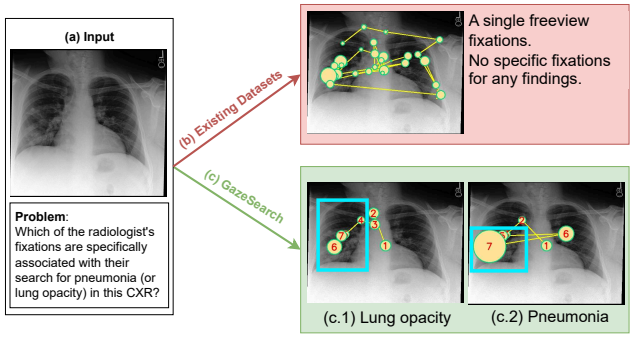

Figure 1. (a) Given a CXR image, we are interested in radiologist's eye movement of radiologist when they search for a finding. (b) But, the existing eye gaze datasets are recorded in a free-view form, where fixations are distributed across the entire CXR image and making it unclear which fixations correspond to specific findings. (c) Our new GazeSearch dataset, where fixation sequence is focused for a specific finding. For example, the gaze sequence in (c.1) targets lung opacity, while (c.2) focuses on pneumonia. Each circle depicts a fixation, with the number and radius indicating its order and duration, respectively.

ing the radiologists' eye movements to improve the transparency and reliability of AI systems in clinical practice [7].

Recognizing the importance of understanding how radiologists' eye movements impact diagnosis, datasets like EGD [30] and REFLACX [5] have been introduced. But, these eye-tracking datasets present two major challenges:

Challenge #1: Free-view format - Existing eye-tracking datasets are collected in a free-view format, where fixations are distributed across the entire CXR image, making it unclear which fixations correspond to specific findings (as shown in Figure 1 (b)). Moreover, these datasets often contain ambiguity and suffer from misalignment between the recorded fixations and the findings in the report, rendering them unsuitable for accurate scan path prediction.

Challenge #2: Lack of finding-aware radiologist's scanpath models - Most existing scanpath prediction models [43,75,77] are designed for general applications and lack the domain-specific expertise needed for radiology. Furthermore, current models trained on medical eye-tracking data are not tailored to the challenges of finding-aware visual search in radiology. For instance, I-AI [51] only associates diseases with abnormalities in specific anatomical areas. While RGRG [60] uses anatomical bounding boxes without considering gaze for report generation.

To address the challenge #1, we propose a finding-aware radiologist's visual search dataset, named **GazeSearch**. Our objective is to minimize the misalignment between the findings extracted from the radiology reports and their corresponding fixations. Insprired by the visual search datasets like COCO-Search18 [75] or Air-D [8], we further process GazeSearch by reducing the fixation length using a radius-based filtering heuristic, ensuring that the direction of fixations remains clear and manageable. Additionally, for every finding, we ensure that the duration of fixations within the location of the given finding is maximized. To create GazeSearch dataset, we utilize the existing free-view eye gaze datasets EGD [30] and REFLACX [5] (Figure 1(b)) to conduct a finding-aware radiologist's visual search dataset (Figure 1(c)), which produces two scanpaths for particular findings e.g., "lung opacity" (Figure 1 (c.1)) and "pneumonia"(Figure 1 (c.2)) in this example. The goal of releasing this dataset is to foster the development of algorithms that better mimic radiologists, especially focusing on understanding observation sequences, attention (duration), frequency on key regions, and expert knowledge [45, 71].

To address challenge #2, we introduce **ChestSearch**, a scanpath prediction architecture that surpasses existing models. ChestSearch builds on a standard meta architecture [13] featuring a feature extractor [24, 39] and a Transformer decoder [62], with two key enhancements. First, we train the feature extractor using the self-supervised MGCA method [66] on the large MIMIC-CXR [29] dataset, providing a strong initialization for training. Second, we utilize the modified cross attention from [12] with a query mechanism to select only relevant fixations for predicting the next fixation. Then, the model's three heads handle distinct tasks: predicting 2D coordinates, duration, and stopping points. Finally, we benchmark ChestSearch against current state-of-the-art visual search models on GazeSearch, showcasing the current advancements in radiology visual search.

Our main contributions are:

- **GazeSearch:** We propose a processing technique that converts free-view eye gaze data into finding-aware radiologist's visual search data. This curated dataset is the first target-present visual search dataset for chest X-ray, making possible deep learning modeling of medical visual search prediction.
- **ChestSearch:** We propose a transformer-based model that utilizes a radiology pretrained feature extractor and query mechanism to choose only relevant fixations to predict subsequent fixations based on previous ones. Additionally, we evaluate ChestSearch against several leading generic scanpath prediction models using our GazeSearch to showcase the current progress in the medical visual search task.

2. Related works

Visual Search Datasets. Search datasets have been rising recently due to the interest in understanding human behavior [8,19,22,28,50,67,80]. This is particularly evident in the general visual domain, where numerous datasets have been created across diverse settings. These datasets cover a wide range of scenarios, from searching for multiple targets simultaneously [22] to focusing on a single or two target categories [19, 80]. Some datasets, like COCO-Search18 [75], feature a large number of target objects, or adopt a Visual Question Answering approach [8]. In contrast, the medical domain has lagged behind in terms of dedicated visual search datasets. Existing medical datasets primarily focus on multi-target search tasks, as demonstrated by datasets like EGD [30] and REFLACX [5]. However, there is a significant lack of search datasets tailored for the medical domain. This paper makes a novel contribution by addressing this research gap. We introduce the first target-present visual search dataset specifically designed for the medical field. This dataset opens up new avenues for research and development in this critical area.

Visual Search Baselines. Parallel to the growth of visual search datasets, significant advancements have been made in scan path prediction accuracy [1, 14, 33, 69, 81]. Early scanpath models mostly rely on sampling fixations from saliency maps [27, 41, 68, 70]. Recent advancements, including the integration of deep neural networks [9, 43, 59, 75, 77, 78], reinforcement learning techniques [9, 75, 77], and transformer-based architectures [10, 43, 53, 76], have significantly deepened our understanding of the temporal

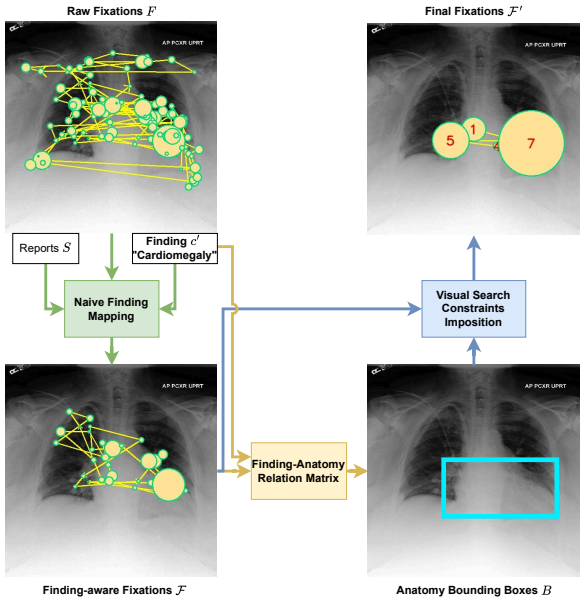

Figure 2. Pipeline of GazeSearch creation, which processes free-view eye gaze data as input and outputs a finding-aware scanpath.

dynamics of human attention. However, generic models are designed for broad application, so the performance of generic visual search models on CXR is uncertain and potentially subpar. This work introduces a transformer-based method that can work well without these restrictive assumptions. Additionally, we further conduct a comparative experiment between state-of-the-art methods from the general visual domain and our proposed method, providing a comprehensive evaluation of their performance in the medical domain.

3. GazeSearch Dataset

When studying free-view eye-tracking datasets from sources like REFLACX [5] and EGD [30], we notice that the eye-tracking data (including both gaze and fixations) is often ambiguous and lacks clarity. This ambiguity comes from the data collection settings, where radiologists look for multiple findings simultaneously. As a result, each fixation captures visual information relevant to multiple findings rather than a specific finding. Therefore, the fixations from these eye-tracking datasets are unsuitable for studying their relationship to specific findings, i.e. addressing the visual search problem. Additionally, when visualizing these gaze points or fixations over an image, they often cover more than 80% of the lung area, even though the actual anomaly area might be much smaller. We calculate the fixation coverage distribution in Supplementary. This raises

Algorithm 1 Radius-based Filtering Procedure

Input: Image width W, image height H, bounding boxes B, max length M, radius r, fixations $\mathcal{F} = \{(x_1, y_1, d_1), (x_2, y_2, d_2), \ldots, (x_n, y_n, d_n)\}$
Output: Filtered fixations $\hat{\mathcal{F}}$
Initialize: $\hat{\mathcal{F}} = (W/2, H/2, 0.3)$
// The last point must be inside B.
$j \leftarrow \max\{i | (x_i, y_i) \in B, (x_i, y_i, d_i) \in \mathcal{F}, 1 \leq i \leq n\}$
// Apply radius heuristic with looping backward.
$c \leftarrow \{(x_j, y_j)\}$, where $(x_j, y_j, d_j) \in \mathcal{F}$
for each point $(x_i, y_i, d_i) \in \mathcal{F}$ from $j-1$ to 1 **do**
  if (x_i, y_i) is within radius r of (x_{i+1}, y_{i+1}) **then**
    $c \leftarrow c \cup \{(x_i, y_i)\}$
  else
    $x \leftarrow \frac{1}{|c|}\sum_k x_k, y \leftarrow \frac{1}{|c|}\sum_k y_k, d \leftarrow \sum_k d_k$,
    where $(x_k, y_k, d_k) \in c$
    $c \leftarrow \{(x_i, y_i)\}$
    $\hat{\mathcal{F}} \leftarrow \hat{\mathcal{F}} \cup \{(x, y, d)\}$
    if $|\hat{\mathcal{F}}| = M$ **then**
      break
    end if
  end if
end for
if $c \neq \{\}$ **and** $|\hat{\mathcal{F}}| < M$ **then**
  $x \leftarrow \frac{1}{|c|}\sum_k x_k, y \leftarrow \frac{1}{|c|}\sum_k y_k, d \leftarrow \sum_k d_k$,
  where $(x_k, y_k, d_k) \in c$
  $\hat{\mathcal{F}} = \hat{\mathcal{F}} \cup \{(x, y, d)\}$
end if

Algorithm 2 Time-spent Constraining Procedure

Input: $\hat{\mathcal{F}} = \{(x_1, y_1, d_1), (x_2, y_2, d_2), \ldots, (x_n, y_n, d_n)\}$, bounding boxes B
Output: Constrained fixations $\mathcal{F}'$
$d^{out} \leftarrow \{\sum_{i=k,(x_i,y_i)\notin B}^{n} d_i | (x_k, y_k, d_k) \in \hat{\mathcal{F}}, 1 \leq k \leq n\}$.
$d^{in} \leftarrow \{\sum_{i=k,(x_i,y_i)\in B}^{n} d_i | (x_k, y_k, d_k) \in \hat{\mathcal{F}}, 1 \leq k \leq n\}$.
$D \leftarrow \{i | d_i^{in} \geq d_i^{out}, 1 \leq i \leq n\}$.
if $1 \notin D$ **then**
  $t \leftarrow \min D$
  $\mathcal{F}' \leftarrow \{(x_i, y_i, d_i) | i \geq t, (x_i, y_i, d_i) \in \hat{\mathcal{F}}\}$
end if

a concern that using the free-view fixations from the given datasets may not be effective and could even pose risks in sensitive sectors like healthcare, particularly for tasks requiring precise localization of specific findings.

To solve this issue, one way is to collect eye-tracking data under the visual search setting directly. However, to collect data by having radiologists examine each of the 14 standard findings in CheXpert [26], would be costly and

time-consuming. Therefore, this paper will propose an alternative technique that leverages eye-tracking data directly from the free-view setting to convert to the finding-aware visual search setting.

Inspired by visual search, we studied the COCO-Search18 [75], Air-D [8], and COCO-Freeview [11,78], and identified two key properties that are required in a visual search dataset:

Property #1: Late fixations tend to converge to more decisive regions of interest (ROIs) [8]. And, Shi et al. [8] have concluded the late fixations are for searching.

Property #2: The fixations within the object of interest tend to have longer durations, while those outside the object are typically shorter.

Based on those two facts, we propose an approach to convert from free-view data into a visual search format, ensuring the filtered fixations retain properties #1 and #2 without sacrificing too many fixations. Figure 2 illustrates the overview of our data processing pipeline, including Naive Finding Mapping (Section 3.1) to clean irrelevant fixations for a given finding, Finding-Anatomy Relation Matrix (Section 3.2) to extract key regions, and finally Visual Search Constraint Imposition (Section 3.3) to produce the fixations that have both visual search properties.

3.1. Naive Finding Mapping

The first problem we must solve is the mismatch between the fixations and the corresponding radiologists' report sentences. The main reason is radiologists observe the images first and then describe their findings, meaning the fixations within the time frame of a sentence may not fully capture the findings reported. Inspired by I-AI [51], we start by completely removing fixations after the current spoken sentence. Let $S = \{s_1, s_2, ..., s_{|S|}\}$ be the sequence of sentences in the transcript. Let $C = \{c_1, c_2, ..., c_m\}$ be the set of possible findings (e.g., CheXpert labels). We define a function $\phi : S \rightarrow C$ where $c_j = \phi(s_i)$ if sentence s_i corresponds to finding c_j. In our implementation, $\phi(\cdot)$ is the Chexbert model [57]. For a target finding $c' \in C$, let $u = \max\{i|\phi(s_i) = c', 1 \leq i \leq |S|\}$. Then, the new finding-aware fixations $\mathcal{F}$ for c' is

$$\mathcal{F} = \{(x_i, y_i, t_i, d_i)|(x_i, y_i, t_i, d_i) \in F, 0 \leq t_i \leq e_u\} \quad (1)$$

where $F = \{(x_1, y_1, t_1, d_1), .., (x_{|F|}, y_{|F|}, t_{|F|}, d_{|F|})\}$ is the free-view fixations, with (x_i, y_i) as spatial coordinates, t_i as captured timestamp, and d_i as duration, and e_u is the ending time of the sentence s_u. From this point onwards, we only use the triplet (x_i, y_i, d_i) and ignore the captured timestamp t_i for our fixation sequence: $\mathcal{F} = \{(x_1, y_1, d_1), \ldots, (x_n, y_n, d_n)\}$, where $n = |\mathcal{F}|$ is the fixation sequence length.

3.2. Finding-Anatomy Relation Matrix

To address this, we leverage the Chest ImaGenome [71] dataset, which offers pairs of findings and their corresponding anatomies, along with anatomy bounding boxes linked to each finding. For precision, we rely on the gold subset of Chest ImaGenome to construct a relation matrix between findings and anatomies. As a final step, a radiologist with over 15 years of experience thoroughly reviews and refines the matrix. The finalized matrix is included in the Supplementary Material. Once the relation matrix is completed, we reference the given finding c' to identify the corresponding anatomies and utilize the ground truth anatomy bounding boxes provided by Chest ImaGenome as our B for the subsequent steps.

3.3. Visual Search Constraint Imposition

After Section 3.1, the maximum fixation sequence length can be over 340 fixations for a finding. Therefore, another task we must solve is reducing this length to an interpretable level for humans.

Utilizing both properties (1) and (2) as our guidance for this process, we perform two main steps: radius-based filtering (to enforce property #1) and time-spent constraining (to enforce property #2). Besides property #1, we observe that the captured fixations from EGD and REFLACX cover one-degree visual angle [5, 30, 38]. Based on that fact, we use the Algorithm 1 to cluster the finding-aware fixations $\mathcal{F}$ to create another fixation set $\hat{\mathcal{F}}$, with a larger radius r of two-degree of visual angle and M is the max length of fixation sequence. Property #1 is enforced by iterating backward from the end to the beginning of the fixation sequence $\mathcal{F}$. Then, we use the Algorithm 2 to make sure the late fixations must spend the most time in the anatomies of interest, which satisfies property #2.

In Algorithms 1 and 2, we define a point (x, y) to be in the bounding box sets B for notation convenience:

$$(x, y) \in B \iff x^{left} \leq x \leq x^{right}, y^{top} \leq y \leq y^{bottom}, \\ \forall (x^{left}, y^{top}, x^{right}, y^{bottom}) \in B \quad (2)$$

To align with the COCO-Search18 dataset, we set the maximum fixation length to $M = 7$ and add a default center as the start fixation. This choice is based on the observation that 95% of the samples in COCO-Search18 have fixation lengths under 7. For the first fixation's duration, we assign 0.3 seconds to it, which reflects the duration of 91% of first fixations in COCO-Search18. In total, GazeSearch has 2,081 images with 413 samples from EGD and 1,668 samples from REFLACX. There are a total of 13 findings. Each sample has fixations for 1 to 6 findings and has a max length of 7, including the default middle fixation. For training and evaluation, we split the dataset into 1,456 samples

Table 1. Usage validation experiments on our GazeSearch. mHC (mean Heatmap Coverage) is the average ratio of the heatmap area to the lung area across all images in GazeSearch.

Method	Fixation Type	AUC	mHC
Naive Classifier	✗	76%	✗
Temporal Classifier	Freeview	81%	91%
	Finding-aware (GazeSearch)	**81%**	**44%**

for training (70%), 208 samples for validation (10%), and 417 samples for testing (20%).

3.4. Usage Validation

Filtering fixations requires discarding information, so it is essential to test and ensure that the new data remains valuable. To validate that GazeSearch's fixations can be as useful as the free-view fixation maps from EGD and REFLACX, we follow Karargyris et al. [30] to perform the Temporal Heatmap experiment. This experiment evaluates whether eye gaze data can enhance classifier performance when using ground truth fixations as temporal inputs. The results, Table 1, indicate that despite using only half the area compared to the free-view setting, performance remains comparable. Detailed implementation of this experiment is provided in the Supplementary.

4. ChestSearch

Given a CXR image I of dimensions $H \times W$ and a target finding c', our objective is to generate a scan-path comprises of fixations $y = \{f_i\}_{i=1}^{n}$, where n represents the number of fixations, and $f_i = (x_i, y_i, d_i)$ is the fixation at 2D coordinate (x_i, y_i) with a duration of d_i.

Figure 3 provides an overview of our method. The process begins by applying a Feature Extractor (Section 4.1) to process I to extract both detailed and high-level visual features. Following this, a Spatiotemporal Embedding (Section 4.2) embeds previous fixations, combined with multi-resolution features, to capture contextual relationships within the sequence. These features are passed through a transformer decoder with cross-attention, self-attention, feedforward layers, and normalization (Section 4.3) to create a decoded latent feature. Finally, the decoded feature is fed into three heads to predict the next fixation: termination prediction (Section 4.4), fixation duration (Section 4.5), and distribution for the next fixation (Section 4.6)

4.1. Feature Extractor

Using features from only the last layer is inadequate for predicting scanpaths [77]. Therefore, we employ ResNet-50 FPN [39] as our Feature Extractor module (FE). Besides, using the ImageNet [16] checkpoint may not be optimal for the medical domain, so we train the FE using a self-supervised approach based on MGCA [66] with the

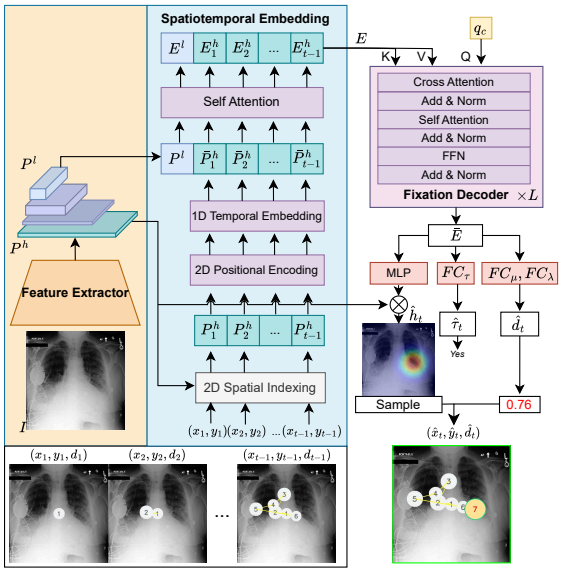

Figure 3. The figure provides a detailed view of ChestSearch. It begins by processing the previous fixations, denoted as $\{(x_i, y_i, d_i)\}_{i=1}^{t-1}$, along with the input chest X-ray image I, through a Feature Extractor and Spatiotemporal Embedding to generates the spatiotemporal embedded feature E. Next, the Fixation Decoder uses a learnable query q_c and the embedded feature E to decode it into a feature $\bar{E}$. From here, three heads use $\bar{E}$ to predict the next fixation coordinates $(\hat{x}_t, \hat{y}_t, \hat{d}_t)$. Here, at step t, the termination head outputs "Yes," indicating that this is the final fixation for the image I.

MIMIC-CXR dataset [29]. From the CXR image I with size $H \times W$, FE produces four multi-scale feature maps $P = \{P^1, \ldots, P^4\}$. Then we need to mimic how human see an image: at first we only see the image at a high level understanding, with no clear details, and then we look carefully to search for what we need [75]. So we use one feature map with low resolution $P^l = P^1 \in \mathbb{R}^{C \times \frac{H}{32} \times \frac{W}{32}}$, where C is the channel dimension, to represent high-level visual feature, and one high-resolution feature map $P^h = P^4 \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$ to represent detailed visual information.

4.2. Spatiotemporal Embedding

Given the previous predicted fixations $\{(x_i, y_i)\}_{i=1}^{t-1}$, P^l, and P^h, we then embed the previous fixations to create the feature list as the input for the decoder in Section 4.3.

2D Spatial Indexing. Every (x_i, y_i), where $0 \leq x_i \leq W$ and $0 \leq y_i \leq H$, is scaled down to the same resolution as of P^h, which result in the new $0 \leq x_i' \leq \frac{W}{4}$ and $0 \leq y_i' \leq \frac{H}{4}$ in our case. Then, we index the feature cell at the coordinate (x_i', y_i') in P^h, called P_i^h. We will have the list of feature $\{P_i^h\}_{i=1}^{t-1}$.

2D Positional Embedding. For every P_i^h, we encode the spatial information by using positional encoding twice, first

in the x-axis, then in the y-axis. As the 2D order is important, we enforce the sinusoid version of positional encoding.

1D Temporal Embedding. We also need to let the model know the order of each fixations. However, the role of fixation order in diagnosing CXR in practice is complicated, so we let the model decide the embedding by applying a learnable position embedding here. This results in the $\{\bar{P}_i^h\}_{i=1}^{t-1}$ sequence of embedded feature.

Self Attention. Finally, we feed $\{\bar{P}_i^h\}_{i=1}^{t-1}$ into several layers of self-attention to aaggregate information so that each position is influenced by the relevant fixations. In the self-attention layers, we also provide the model with a low-resolution feature map P^l to supply high-level feature information. This intuition is also proven effected empirically, as it will be shown later in Section 5.4. The final embeddings are $E = \{E^l\} \cup \{E_i^h\}_{i=1}^{t-1}$, where $E^l \in \mathbb{R}^{D \times \frac{H}{32} * \frac{W}{32}}$ and $E_i^h \in \mathbb{R}^D$.

4.3. Fixation Decoder

At this layer, we have the finding list $q = \{q_c\}_c^{|q|}$ which serves as the set of queries. The number of queries is the number of findings in our dataset $|q| = 13$ with $q_c \in \mathbb{R}^D$ is a learnable embedding for the current finding query c. The previous module (Section 4.2) gives us the embeddings of previous fixations E.

The Fixation Decoder module is the modified transformer decoder [12] including the blocks as shown in Figure 3. The cross-attention block uses the query embedding q as the query input Q, with E serving as both key (K) and value (V). This allows the model to capture the correlations among previous fixations and accurately predict the next fixation. The resulting feature then passes through self-attention layers, residual connections, normalization, and a feed-forward network. This process repeats for L layers in the decoder. The final output $\bar{E} \in \mathbb{R}^{|q| \times D}$ is then processed by three different heads.

4.4. Termination Head

A fixation sequence's length can vary, so our model needs to learn when to stop. To achieve this, we use a head consisting of a fully connected (FC) layer followed by a sigmoid function that maps $\bar{E}$ to termination value i.e., $\hat{\tau}_t^c \in \mathbb{R} = \text{sigmoid}(FC_\tau(\bar{E}))$.

4.5. Duration Head

The duration can be considered as a Gaussian distribution. We use $\bar{E}$, then regress it into a mean value $\mu_{d_t} = FC_\mu(\bar{E})$ and a log-variance $\lambda_{d_t} = FC_\lambda(\bar{E})$:

$$\begin{aligned} \hat{d}_t &= \mu_{d_t} + \epsilon_{d_t} \cdot \exp(0.5\lambda_{d_t}), \\ \epsilon_{d_t} &\sim \mathcal{N}(0,1) \end{aligned} \tag{3}$$

where ϵ_{d_t} noise gives our prediction a probabilistic characteristic, and $\hat{d}_t \in \mathbb{R}^{|q|}$ is the duration prediction. The inspiration comes from using the reparameterization trick [18], which allows us to backpropagate from the label back to the normal distribution.

4.6. Distribution Head

Because fixation is random in nature, we predict a 2D distribution in the form of a heatmap $\hat{h}_t \in [0,1]^{|q| \times (\frac{H}{4} * \frac{W}{4})}$. Formally, we compute:

$$\begin{aligned} \bar{E}' &= \text{MLP}(\bar{E}) \\ \hat{h_t} &= \text{sigmoid}(\text{Matmul}(\bar{E}', P^h)) \end{aligned} \tag{4}$$

where $\text{Matmul}(\cdot,\cdot)$ is the matrix multiplication between two input tensors, and $\bar{E}' \in \mathbb{R}^{|q| \times D}$ is the latent embedding prepared for heatmap generation. At inference, we sample the next 2D coordinate $\hat{f}_t = (\hat{x}_t, \hat{y}_t)$ from the distribution map $\hat{h}_t$ for every given timestamp t.

4.7. Objective Functions

ChestSearch has three objectives, each corresponding to one of its heads: the loss between the ground truth and predicted distributions, the loss for termination, and the loss for duration.

The termination loss is just a standard binary cross-entropy between the predicted termination value $\hat{\tau}_t$ and the corresponding ground truth τ_t.

$$\mathcal{L}_\tau = -\tau_t \log(\hat{\tau}_t) - (1-\tau_t)\log(1-\hat{\tau}_t), \tag{5}$$

The distribution loss is defined as focal pixel-wise loss:

$$\mathcal{L}_h = -\frac{1}{N}\sum_{ij} \begin{cases} (1-\hat{h}_{ij})^\gamma \log(\hat{h}_{ij}) & \text{if } h_{ij} = 1, \\ (1-h_{ij})^\alpha (\hat{h}_{ij})^\gamma \log(1-\hat{h}_{ij}) & \text{otherwise,} \end{cases} \tag{6}$$

where $0 \le i \le \frac{H}{4}$, $0 \le j \le \frac{W}{4}$ are the 2D indexes, $N = \frac{H}{4} * \frac{W}{4}$ is the number of values, α and γ are the hyper-parameters indicating the importance of each pixel. The duration loss is defined as the L1 loss, i.e., $\mathcal{L}_d = |\hat{d}_t - d_t|$.

Finally, we train all three losses jointly.

$$\mathcal{L} = \mathcal{L}_\tau + \mathcal{L}_h + \mathcal{L}_d \tag{7}$$

5. Experiments

5.1. Implementation and Metrics

Implementation details. All images are scaled down to 224×224 for computing efficiency. The Fixation Decoder has $L = 6$ layers with a hidden dimension $D = 384$. The MLP of Fixation Distribution Head consists of 384 units with 3 layers and ReLU activation. Eq. (6) has $\alpha = 4$ $\gamma = 2$ based on the best validation results. The Feature Extractor's backbone is ResNet-50 [24], and we obtain the ResNet-50 checkpoint using MGCA [66] for 50 epochs with a batch

Table 2. Performance comparison between our ChestSearch and SOTA visual search methods.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓	STDE ↑
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration		
IRL [75]	0.1495	-	0.8248	0.6402	0.7688	0.6998	-	6.6250	0.7043
FFMs [77]	0.2766	-	0.8914	0.6567	0.8785	0.8140	-	5.9221	0.8055
ChenLSTM [9]	0.2751	0.2153	0.8825	0.6222	0.8731	0.7940	0.6384	5.3468	0.7841
ChenLSTM-ISP [10]	0.2863	0.2205	0.8847	0.6430	0.8721	0.7980	0.6504	5.2895	0.7865
Gazeformer [43]	0.2971	0.2042	0.9080	0.6506	0.9035	0.8147	0.5901	5.1024	0.8030
Gazeformer-ISP [10]	0.2736	0.2146	0.9038	0.6181	0.8892	0.8031	0.6755	5.1905	0.7875
HAT [76]	0.3120	-	0.9064	0.6443	0.9065	0.8138	-	5.0613	0.8006
Our ChestSearch	**0.3321**	**0.2232**	**0.9173**	**0.6790**	**0.9174**	**0.8293**	**0.6951**	**4.8831**	**0.8089**

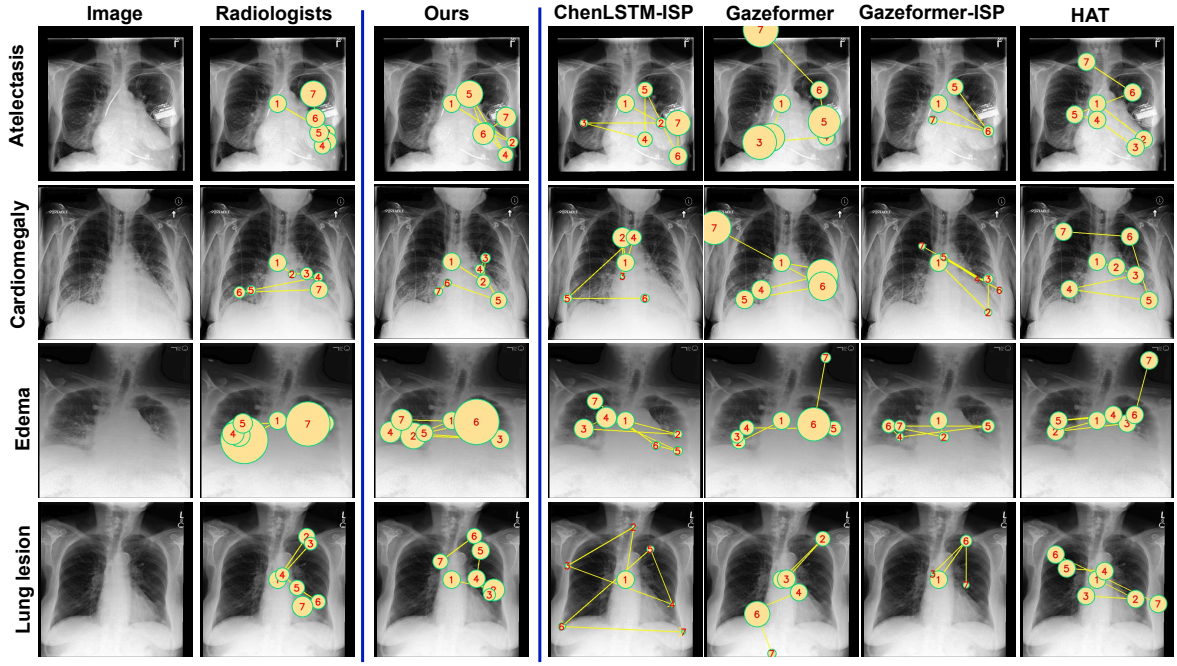

Figure 4. Qualitative results between our ChestSearch compared with ChenLSTM-ISP, Gazeformer, Gazeformer-ISP, and HAT. Four different findings (rows) including Atelectasis, Cardiomegaly, Edema, and Lung lesion are shown from the top to bottom. Each circle represents a fixation, with the number and radius indicating its order and duration, respectively. As HAT only predicts 2D coordinates, we let all predicted fixations of HAT have the same radius.

size of 144. We then finetune this checkpoint jointly with the full pipeline. We train the full pipeline for 30,000 iterations with a learning rate of 1×10^{-5} and a batch size of 32. The entire training process was conducted using AdamW [40], on a single A6000 GPU with 48GB of RAM.

Metrics. We evaluate fixation scanpath prediction accuracy using various metrics: ScanMatch [15, 58] applies the Needleman-Wunsch algorithm [44] to compare fixation locations and durations; MultiMatch [17] assesses similarity across five dimensions; String-Edit Distance (SED) [6, 20] compares character strings representing image regions; and Scaled Time-Delay Embedding (STDE) [68] measures mean minimum Euclidean distances between subsequences of predicted and ground truth scanpaths.

Compared Methods. We evaluate several state-of-the-art (SOTA) visual search methods on our GazeSearch: IRL [75], FFMs [77], ChenLSTM [9], Gazeformer [43], ChenLSTM-ISP [10], Gazeformer-ISP [10], and HAT [76]. Note that Gazeformer and Gazeformer-ISP require a pre-trained CLIP component to encode the finding names, so we replace its default CLIP with BiomedCLIP [82]. We adhere to the original training practices for all baselines. For more details, please refer to the Supplementary.

5.2. Quantitative results

Table 2 demonstrates the proposed method's superior performance, surpassing SOTA approaches. Note that IRL, FFMs, and HAT do not predict fixation duration, so their evaluation on this metric is excluded. IRL and FFMs face challenges with sample efficiency due to re-

inforcement learning pipelines, while ChenLSTM variants and ISP methods are limited by their specialized modules—ChenLSTM relies on pretrained object detectors and ISPs on Observer-Centric modules. HAT and Gazeformer overgeneralize and fail to fully leverage domain-specific information by design, with HAT ignoring duration data and Gazeformer relying heavily on CLIP for zero-shot visual search. Our method avoids these limitations. High scores in metrics such as ScanMatch, MultiMatch, SED, and STDE demonstrate our method's capability to effectively capture complex scanpath dynamics, setting a new standard in chest X-ray target-present visual search.

5.3. Qualitative results

Figure 4 presents a qualitative comparison of scanpath patterns across different radiology findings and models, including radiologists and several state-of-the-art approaches. Generally, ChestSearch predicts more consistent and radiologist-like fixations than other methods. ChenLSTM-ISP often exhibits scattered, less focused patterns, while Gazeformer-ISP may overlook key areas or focus on fewer locations. Although Gazeformer aligns better with ground truth than its ISP variant, it occasionally misses critical regions, such as lung lesions. HAT performs reasonably well but frequently covers the entire lung, even when attention should be limited to smaller areas, such as in cardiomegaly. In contrast, our ChestSearch shows fixation patterns more closely resembling those of radiologists, outperforming other state-of-the-art methods. Overall, Figure 4 underscores the effectiveness of our approach in mimicking expert gaze patterns across different findings. Additional comparison will be included in the Supplementary.

5.4. Ablation study

To study the design choice of our proposed architecture, we ablate our method under several aspects.

The importance of low- and high-resolution feature maps. In Section 4.2, guided by our intuition, we use two feature maps: a low-resolution map for high-level visual understanding and a high-resolution map for detailed visual understanding. These are concatenated into a single tensor for the Self-Attention layer, with the low-resolution feature serving as a *reference* and the high-resolution feature *indexed* using 2D Spatial Indexing to generate temporal features. Ablation results in Table 3 show that omitting 2D Spatial Indexing results in a significant performance drop due to the loss of temporal information. Conversely, not using the reference feature before Self-Attention has a lesser impact. The optimal performance is achieved by using the low-resolution feature as the reference and the high-resolution feature for 2D indexing, aligning with our intuitive design choices.

Initial Feature Extractor's weight contribution. This ablation studies the effect of the initial weight for the Feature Extractor(Section 4.1), shown in Table 4. In conclusion, using ImageNet checkpoint can give a decent performance. But with a better checkpoint, the performance is higher. This shows the robustness of our architecture.

Table 3. The role of low- and high-resolution feature maps.

Reference	Indexing	ScanMatch ↑		MultiMatch ↑	SED ↓	STDE ↑
		w/o Dur.	w/ Dur.			
P^l	✗	0.1848	0.2029	0.7070	6.3636	0.7066
P^h	✗	0.1939	0.1925	0.7058	6.1424	0.7184
✗	P^l	0.3077	0.2177	0.7927	5.0180	0.8027
✗	P^h	0.3176	0.2204	0.7985	4.9078	0.8035
P^l	P^l	0.3129	0.2228	0.7989	4.9100	0.8060
P^h	P^h	0.3221	0.2229	0.8015	5.0277	0.8058
P^h	P^l	0.3184	0.2210	0.8022	5.0224	0.8057
P^l	P^h	**0.3321**	**0.2232**	**0.8076**	**4.8831**	**0.8089**

Table 4. Ablation study of choosing initial weight.

Inital Weight	ScanMatch ↑		MultiMatch ↑	SED ↓	STDE ↑
	w/o Dur.	w/ Dur.			
Random Init.	0.3130	0.2205	0.79224	5.0331	0.8058
ImageNet	0.3238	0.2224	0.79942	4.9723	0.8081
Ours (Self-supervised)	**0.3321**	**0.2232**	**0.80762**	**4.8831**	**0.8089**

6. Conclusion

This paper addresses two key challenges: ambiguous fixations in existing eye-tracking datasets and the absence of a finding-aware radiologist's scanpath model. Drawing inspiration from visual search datasets in general domains, we align findings with fixations, manage fixation durations using a radius-based heuristic, and constrain fixations on duration to produce the first finding-aware visual search dataset, GazeSearch. Our dataset reflects two key properties of visual search behavior: #1 late fixations tend to converge on decisive regions of interest, and #2 fixations within objects of interest are typically longer in duration compared to those outside. We then propose ChestSearch that utilizes self-supervised training to obtain a medical pretrained feature extractor and a query mechanism to select relevant fixations for predicting subsequent ones. The extensive benchmark shows ChestSearch 's ability to generate radiologist-like scanpaths, serving as a strong baseline for future research.

Discussion: Our work impacts the behavioral vision literature in the medical domain, where (i) modeling and replicating radiologists' behavior has not been explored, (ii) understanding the understanding of finding-aware visual search and their integration with Deep Learning remains poorly understood [45]. These are critical for advancing diagnostics in radiology, enhancing decision-making processes, and enabling the future development of collaborative interactions between radiologists and AI systems.

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.

References

[1] Hossein Adeli and Gregory Zelinsky. Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *CVPR Workshops*, 2018. 2

[2] Akash Awasthi, Ngan Le, Zhigang Deng, Rishi Agrawal, Carol C Wu, and Hien Van Nguyen. Bridging human and machine intelligence: Reverse-engineering radiologist intentions for clinical trust and adoption. *Computational and Structural Biotechnology Journal*, 2024. 1

[3] Atallah Baydoun et al. Artificial intelligence applications in prostate cancer. *Prostate cancer and prostatic diseases*, 27(1):37–45, 2024. 1

[4] Sebastien Benzekry. Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clinical Pharmacology & Therapeutics*, 108(3):471–486, 2020. 1

[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. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022. 2, 3, 4

[6] Stephan A Brandt and Lawrence W Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of cognitive neuroscience*, 9(1):27–38, 1997. 7

[7] Nora Castner, Lubaina Arsiwala-Scheppach, Sarah Mertens, Joachim Krois, Enkeleda Thaqi, Enkelejda Kasneci, Siegfried Wahl, and Falk Schwendicke. Expert gaze as a usability indicator of medical ai decision support systems: a preliminary study. *NPJ Digital Medicine*, 7(1):199, 2024. 2

[8] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[9] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 7

[10] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024. 2, 7

[11] Yupei Chen et al. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5031–5040, 2022. 4

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 6

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 2

[14] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing (IEEE TIP)*, 2018. 2

[15] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42:692–700, 2010. 7

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[17] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44:1079–1100, 2012. 7

[18] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 6

[19] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 2

[20] Tom Foulsham and Geoffrey Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*, 8(2):6–6, 2008. 7

[21] Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discover Artificial Intelligence*, 4(1):15, 2024. 1

[22] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015. 2

[23] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024. 1

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv e-prints. *arXiv preprint arXiv:1512.03385*, 10, 2015. 2, 6

[25] Nehmat Houssami, Georgia Kirkpatrick-Jones, Naomi Noguchi, and Christoph I Lee. Artificial intelligence (ai) for the early detection of breast cancer: a scoping review to assess ai's potential in breast screening practice. *Expert review of medical devices*, 16(5):351–362, 2019. 1

[26] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 3

[27] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 1998. 2

[28] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[29] Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 2019. 2, 5

[30] Alexandros Karargyris et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*, 8(1):1–18, 2021. 2, 3, 4, 5

[31] Enkelejda Kasneci et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023. 1

[32] Khadija Khaldi, Vuong D Nguyen, Pranav Mantini, and Shishir Shah. Unsupervised person re-identification in aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 260–269, 2024. 1

[33] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 2

[34] Minh-Quan Le, Alexandros Graikos, Srikar Yellapragada, Rajarsi Gupta, Joel Saltz, and Dimitris Samaras. ∞-brush: Controllable large image synthesis with diffusion models in infinite dimensions. *arXiv preprint arXiv:2407.14709*, 2024. 1

[35] Minh-Quan Le, Tam V Nguyen, Trung-Nghia Le, Thanh-Toan Do, Minh N Do, and Minh-Triet Tran. Maskdiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2874–2881, 2024. 1

[36] Ngan Le, Vidhiwar Singh Rathour, Kashu Yamazaki, Khoa Luu, and Marios Savvides. Deep reinforcement learning in computer vision: a comprehensive survey. *Artificial Intelligence Review*, pages 1–87, 2022. 1

[37] Ngan Le, James Sorensen, Toan Bui, Arabinda Choudhary, Khoa Luu, and Hien Nguyen. Enhance portable radiograph for fast and high accurate covid-19 monitoring. *Diagnostics*, 11(6):1080, 2021. 1

[38] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, 45(1), 2013. 4

[39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 5

[40] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[41] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015. 2

[42] Mohammad Muzaffar Mir, Gulzar Muzaffar Mir, Nadeem Tufail Raina, Saba Muzaffar Mir, Sadaf Muzaffar Mir, Elhadi Miskeen, Muffarah Hamid Alharthi, and Mohannad Mohammad S Alamri. Application of artificial intelligence in medical education: current scenario and future perspectives. *Journal of advances in medical education & professionalism*, 11(3):133, 2023. 1

[43] Sounak Mondal et al. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7

[44] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 7

[45] José Neves, Chihcheng Hsieh, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Anderson Maciel, Andrew Duchowski, Joaquim Jorge, and Catarina Moreira. Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. *European Journal of Radiology*, page 111341, 2024. 2, 8

[46] E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael Ryoo. Instance-aware generalized referring expression segmentation, 2024. 1

[47] Vuong D Nguyen, Khadija Khaldi, Dung Nguyen, Pranav Mantini, and Shishir Shah. Contrastive viewpoint-aware shape learning for long-term person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1041–1049, 2024. 1

[48] Vuong D Nguyen, Samiha Mirza, Abdollah Zakeri, Ayush Gupta, Khadija Khaldi, Rahma Aloui, Pranav Mantini, Shishir K Shah, and Fatima Merchant. Tackling domain shifts in person re-identification: A survey and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4149–4159, 2024. 1

[49] Hai Nguyen-Truong, E-Ro Nguyen, Tuan-Anh Vu, Minh-Triet Tran, Binh-Son Hua, and Sai-Kit Yeung. Vision-aware text features in referring image segmentation: From object understanding to context understanding, 2024. 1

[50] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014. 2

[51] Trong Thang Pham, Jacob Brecheisen, Anh Nguyen, Hien Nguyen, and Ngan Le. 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*, pages 7850–7859, 2024. 1, 2, 4

[52] Trong Thang Pham, Ngoc-Vuong Ho, Nhat-Tan Bui, Thinh Phan, Patel Brijesh, Donald Adjeroh, Gianfranco Doretto, Anh Nguyen, Carol C. Wu, Hien Nguyen, and Ngan Le. Fg-cxr: A radiologist-aligned gaze dataset for enhancing interpretability in chest x-ray report generation. *ACCV*, 2024. 1

[53] Mengyu Qiu, Yi Guo, Mingguang Zhang, Jingwei Zhang, Tian Lan, and Zhilin Liu. Simulating human visual system based on vision transformer. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, 2023. 2

[54] Daniel L Rubin. Artificial intelligence in imaging: the radiologist's role. *Journal of the American College of Radiology*, 16(9):1309–1317, 2019. 1

[55] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. 1

[56] Yuping Shang, Silu Zhou, Delin Zhuang, Justyna Żywiołek, and Hasan Dincer. The impact of artificial intelligence application on enterprise environmental performance: Evidence from microenterprises. *Gondwana Research*, 131:181–195, 2024. 1

[57] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020. 4

[58] Hiroyuki Sogo. Gazeparser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior research methods*, 45:684–695, 2013. 7

[59] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using IOR-ROI recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019. 2

[60] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *CVPR*, 2023. 2

[61] Kim Hoang Tran, Phuc Vuong Do, Ngoc Quoc Ly, and Ngan Le. Unifying global and local scene entities modelling for precise action spotting. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. 1

[62] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[63] Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *International Journal of Computer Vision*, 131(1):302–323, 2023. 1

[64] Khoa Vo, Kashu Yamazaki, Phong X Nguyen, Phat Nguyen, Khoa Luu, and Ngan Le. Contextual explainable video representation: Human perception-based understanding. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 1326–1333. IEEE, 2022. 1

[65] Stephen Waite et al. A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era of artificial intelligence. *Academic Radiology*, 27(1):26–38, 2020. 1

[66] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 2, 5, 6

[67] Shuo Wang et al. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 2015. 2

[68] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2, 7

[69] Zijun Wei, Hossein Adeli, Minh Hoai, Gregory Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predict visual attention. In *NeurIPS*, 2016. 2

[70] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[71] 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. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. 2, 4

[72] Nan Wu et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019. 1

[73] Kashu Yamazaki, Khoa Vo, Quang Sang Truong, Bhiksha Raj, and Ngan Le. Vltint: Visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In *Proceedings of the AAAI Conference on Artificial intelligence*, volume 37, pages 3081–3090, 2023. 1

[74] Kashu Yamazaki, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain Sciences*, 12(7):863, 2022. 1

[75] Zhibo Yang et al. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5, 7

[76] Zhibo Yang et al. Unifying top-down and bottom-up scanpath prediction using transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 2, 7

[77] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 7

[78] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*, 2023. 2, 4

[79] Nurullah Yüksel, Hüseyin Rıza Börklü, Hüseyin Kürşad Sezer, and Olcay Ersel Canyurt. Review of artificial intelligence applications in engineering design perspective. *Engineering Applications of Artificial Intelligence*, 118:105697, 2023. 1

[80] Gregory Zelinsky et al. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[81] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018. 2

[82] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 7