# GazeSearch: Radiology Findings Search Benchmark Supplementary Materials

## 1. Background Concepts

**Fixations** are the periods (points) when the eyes remain relatively still, focusing on a point in the visual field to gather detailed information. They provide insights into cognitive processing, with longer fixations indicating areas of interest or complexity requiring more attention [4].

**Visual angle** refers to the dimensions of retinal features described in terms of the projected size of a scene, measured in degrees. From this degree (e.g. one degree visual angle), we can convert it to pixels. In EGD [7] and REFLACX [1], visual angle is given under pixel units.

## 2. Fixation Coverage Distribution

By calculating the distribution of area ratio based on the lung area versus the covered fixation heatmap, we realize that this observation happens on most samples in the current radiology eye-tracking datasets (EGD and REFLACX), as shown in Figure 1a(a). After Section 3 in the main paper, we again plot the distribution, Figure 1b, to show that we scale down the covered area down to mostly 35-60%, indicating that the filtered fixations are focusing on more specific areas.

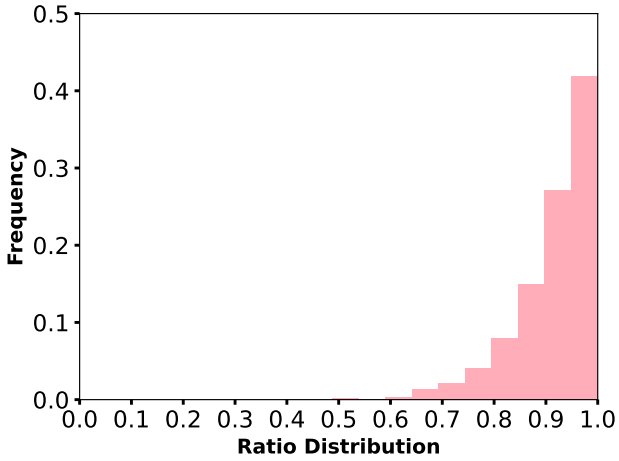
## 3. Details of Finding-Anatomy Relation Matrix

The matrix details are shown in Table 1. For every anatomy in the matrix, we can get several important information: its bounding box, the findings associated with it, and the radiology report about it. In our work, we use this relationship to get all bounding boxes for every finding.

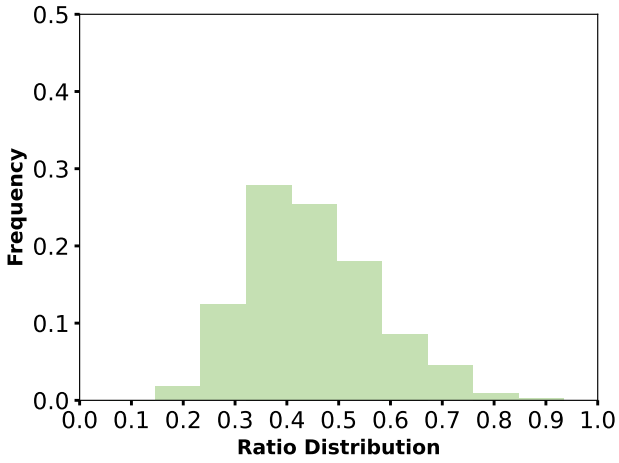
## 4. Addtional Details for GazeSearch's Usage Validation

### 4.1. Temporal Classifier

The Temporal Classifier [7] is designed to integrate both image data and temporal fixation heatmaps for classification purposes, providing a richer context than using image data alone. Figure 2 shows the architecture of Temporal Classifier, including Image Encoding: step (blue), Fixation Heatmap Encoding (green), and Classifying (orange).




(a) Heatmap area ratio distribution of free-viewing data.




(b) Heatmap area ratio distribution of filtered data.

Figure 1. Histograms of heatmap area ratios. The frequency indicates how many times a particular ratio appears. The ratio itself is calculated by dividing the area of the heatmap within the lung area(ratio = heatmap area/lung area). The majority of attention heatmaps from free-view fixations cover more than 80% of the lung area (a). In contrast, the covered area in (b) is smaller compared to (a), but it is not drastically reduced.

*Input:* The image $I^i$ is accompanied by a sequence of $n$ temporal fixation heatmaps, $h^i = h^i_{k\,k=1}{}^n$. Each heatmap

Table 1. Finding-Anatomy Relation Matrix Details. For every anatomy (row), it may (Yes) or may not (No) contains some clues regarding the radiology findings (columns). For example, "cardiomegaly" can only be seen if we look at "cardiac silhouette" or "mediastinum" regions. The radiology findings are the 13 abnormal findings from CheXpert [6]. The anatomies are the 27 common anatomies from Chest ImaGenome [13].

| Anatomies | enlarged cardiomediastinum | cardiomegaly | lung opacity | lung lesion | edema | consolidation | pneumonia | atelectasis | pneumothorax | pleural effusion | pleural other | fracture | support devices |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aortic arch | Yes | No | No | No | No | No | No | No | No | No | No | No | No |
| cardiac silhouette | Yes | Yes | No | No | yes | No | No | No | No | no | No | No | Yes |
| carina | No | No | No | No | No | No | No | No | No | No | No | No | Yes |
| mediastinum | Yes | Yes | no | no | yes | no | no | no | no | no | No | no | Yes |
| upper mediastinum | Yes | no | no | no | yes | no | No | No | no | no | No | No | Yes |
| cavoatrial junction | No | No | No | No | No | No | No | No | No | No | No | No | Yes |
| trachea | yes | No | No | no | No | No | No | No | No | no | No | No | Yes |
| left apical zone | No | No | Yes | Yes | Yes | Yes | yes | No | Yes | Yes | Yes | no | Yes |
| left clavicle | No | No | No | no | No | No | No | No | no | No | No | Yes | no |
| left costophrenic angle | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | yes |
| left hemidiaphragm | No | No | Yes | Yes | Yes | yes | yes | Yes | yes | Yes | yes | No | Yes |
| left hilar structures | No | No | Yes | Yes | Yes | Yes | Yes | no | no | no | No | No | Yes |
| left lower lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | Yes |
| left lung | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | Yes |
| left mid lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | yes | Yes | Yes | No | yes |
| left upper lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | yes | No | Yes |
| right apical zone | No | No | Yes | Yes | Yes | Yes | yes | no | Yes | Yes | Yes | no | Yes |
| right atrium | No | No | No | No | No | No | No | No | No | No | No | No | Yes |
| right clavicle | No | No | No | no | No | No | No | No | no | No | No | Yes | no |
| right costophrenic angle | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | Yes |
| right hemidiaphragm | No | No | Yes | Yes | yes | Yes | Yes | Yes | yes | Yes | yes | No | Yes |
| right hilar structures | No | No | Yes | Yes | Yes | Yes | Yes | no | no | no | No | No | Yes |
| right lower lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | Yes |
| right lung | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | no | Yes |
| right mid lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | yes | Yes | yes | no | yes |
| right upper lung zone | No | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | yes | No | yes |
| svc | No | No | No | No | No | No | No | No | no | No | No | No | Yes |

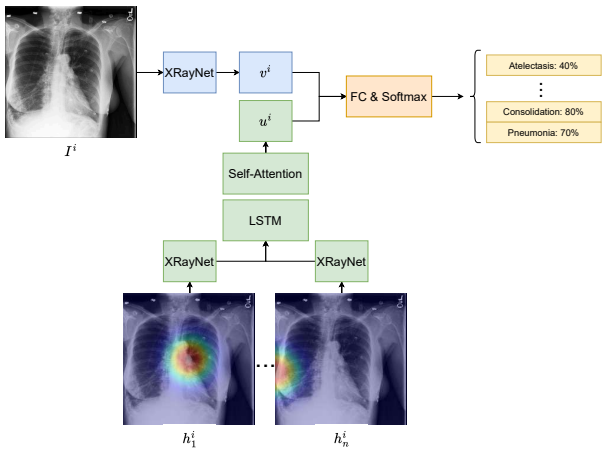



Figure 2. The architecture of Temporal Classifier (Section 4.1).

$h_k^i$ is generated by applying a Gaussian kernel with a sigma equivalent to two degrees of visual angle to the corresponding fixation point $f_k^i = (x_k, y_k)$. For further details, refer to [7, 9].

*Image Encoding:* The image $I$ is processed through XRayNet [7], a 64-channel convolutional layer (kernel size 7, stride 2), max-pooling, batch normalization, and a fully-connected layer with 128 units, resulting in a fixed vector representation of the image, denoted $v^i$.

*Fixation Heatmap Encoding:* Each heatmap $h_k^i$ is similarly processed by a shared instance of XRayNet, tailored specifically for heatmap encoding. These encoded heatmap features are then passed through a one-layer bidirectional LSTM equipped with a self-attention mechanism to summarize the temporal sequence, resulting in a representation denoted $u^i$.

*Classifying:* Now, both the image and heatmap representations are concatenated as $[v^i, u^i]$, and this combined data flows into the final classification layer.

*The Naive Classifier* is Temporal Classifier, but without the Fixation Heatmap Encoding step. In other words, the Naive Classifier is a simple CNN classifier that only uses CXR $I$ as the input.

### 4.2. Implementation Details

*Classification Data:* We structure GazeSearch as a multi-label classification dataset. Given a sample, the model predicts whether a specific finding exists or not.

*Training:* We strictly follow Karargyris et al. [7] to use the Adam [8] optimizer, with an initial learning rate set to 0.001, adjusted via a triangular schedule with fixed decay, a batch size of 16, and a dropout rate of 0.5.

The results, as seen in the main paper's Table 1, agree with the conclusion from the original paper [7], that it highlights the importance of incorporating temporal free-view fixation data. However, the mean heatmap coverage (mHC) of the temporal free-view fixations goes above 90% of the lung area. In contrast, our data processing pipeline (main paper's Section 3) reduces coverage by more than half, but the benefit of temporal information remains the same, suggesting its robustness.

## 5. Additional Details for Comapred Methods

This section provides more technical details for Section 5.1 in the main paper. IRL [14], FFMs [16], and HAT [15] are trained from scratch using their published implemen-

Table 2. Ablation study of selecting Feature Extractor feature maps as low and high resolution.

| Method | | ScanMatch ↑ | | MultiMatch ↑ | SED ↓ | STDE ↑ |
|---|---|---|---|---|---|---|
| $l$ | $h$ | w/o Dur. | w/ Dur. | | | |
| 1 | 3 | 0.3295 | 0.2219 | 0.79445 | 4.8997 | 0.8076 |
| 1 | 2 | 0.3220 | 0.2205 | 0.7886 | 5.0122 | 0.8065 |
| 2 | 4 | 0.3302 | 0.2230 | 0.79545 | 4.8945 | 0.8077 |
| 2 | 3 | 0.3240 | 0.2206 | 0.78835 | 5.0089 | 0.8061 |
| 3 | 4 | 0.3280 | 0.2212 | 0.79105 | 4.9875 | 0.8066 |
| 1 | 4 | **0.3321** | **0.2232** | **0.79815** | **4.8831** | **0.8089** |

tation. However, Gazeformer [11], GazeformerISP [3], ChenLSTM [2], and ChenLSTM-ISP [3] need to be modified in order to work in medical domain.

- Gazeformer and Gazeformer-ISP use pretrained Resnet-50 [5] and RoBERTa [10] as their visual and text encoder. As these modules are from general domain, we must change them to work properly. In this case, we use BiomedCLIP [17], trained on more than 15 million pair medical image - text, as it has both visual encoder and text encoder.
- ChenLSTM and ChenLSTM-ISP require a VQA Model. According to [2,3], this VQA module is equivalent to object detection module for visual search task. Specifically, Chen et al. [2, 3] use CenterNet [18] for both architectures. Because EGD and REFLACX do not provide annotation bounding boxes for all findings. We use the pretrained CenterNet published in the VinBigData Chest X-ray Abnormalities Detection challenge [12] (0.30 mAP@0.5). Note that, a 0.30 mAP@0.5 represents the current state-of-the-art for VinBigData Chest X-ray Abnormalities Detection dataset. For CheXpert findings that are not detected, CenterNet is configured to predict the bounding box for the entire lung.

Finally, we follow the published implementation from each method's original article.

## 6. Contribution of different Feature Extractor levels

In Section 4 of our main paper, we choose the low resolution feature map $P^l$ to be $P^1$ ($l = 1$) and high resolution feature map $P^h$ ($h = 4$) to be $P^h$ based on the GazeSearch's validation results. We notice that as $l$ and $h$ are further apart, the performance slightly increase. Table 2 shows all combinations of $l$ and $h$ from 1 to 4 on the test set, demonstrating a similar trend.

## 7. Additional Qualitative Results

In this section, we extend our comparison of ChestSearch to various state-of-the-art (SOTA) methods across different findings (Figure 3), different SOTAs from the main paper (Figure 4), and other chest X-ray (CXR) images from our test set (Figure 5). Overall, ChestSearch demonstrates the ability to learn and predict scanpaths similar to those of radiologists, outperforming other SOTA models from the general domain.

Figure 3 compares ChestSearch with ChenLSTM-ISP, Gazeformer, Gazeformer-ISP, and HAT on the remaining findings (Figure 4 in our main paper provides 4 findings). To give more details, here are the radiology reports:

- First row (Enlarged cardiomediastinum): "A large dilated, debris-filled, possibly fluid-filled esophagus is again appreciated, abutting the right mediastinum, in this patient with known achalasia."
- Second row (Fracture): "Multiple remote right-sided rib fractures are again noted."
- Third row (Lung opacity): "Frontal and lateral views of the chest demonstrate low lung volumes and bibasilar opacities. Bibasilar opacities in the setting of low lung volumes could represent atelectasis, but multifocal infection is also a possibility."
- Fourth row (Pleural effusion): "Moderate bilateral pleural effusions, stable on the right, decreased on the left."
- Fifth row (Pleural other): "Left lateral basilar pleural thickening is unchanged."
- Sixth row (Pneumonia): "Multiple focal patchy opacities are seen in the bilateral lungs, concerning for multifocal pneumonia."
- Seventh row (Pneumothorax): " Right pneumothorax post surgery with three chest tubes in place."
- Eighth row (Consolidation): "There is persistent airspace consolidation in the right mid lung and right medial lung base which may reflect residual pneumonia/aspiration."
- Nineth row (Support devices): "Enteric tube is now seen with side-port just past the GE junction. Endotracheal tube tip is 4.9 cm from the carina. "

Figure 4 compares ChestSearch with IRL and FFMs on the same samples as in Figure 4 of our main paper. Here are the radiology reports:

- First row (Atelectasis): "Patchy left basilar opacity suggests minor unchanged atelectasis."
- Second row (Cardiomegaly): "Mediastinum: Stable cardiomegaly."

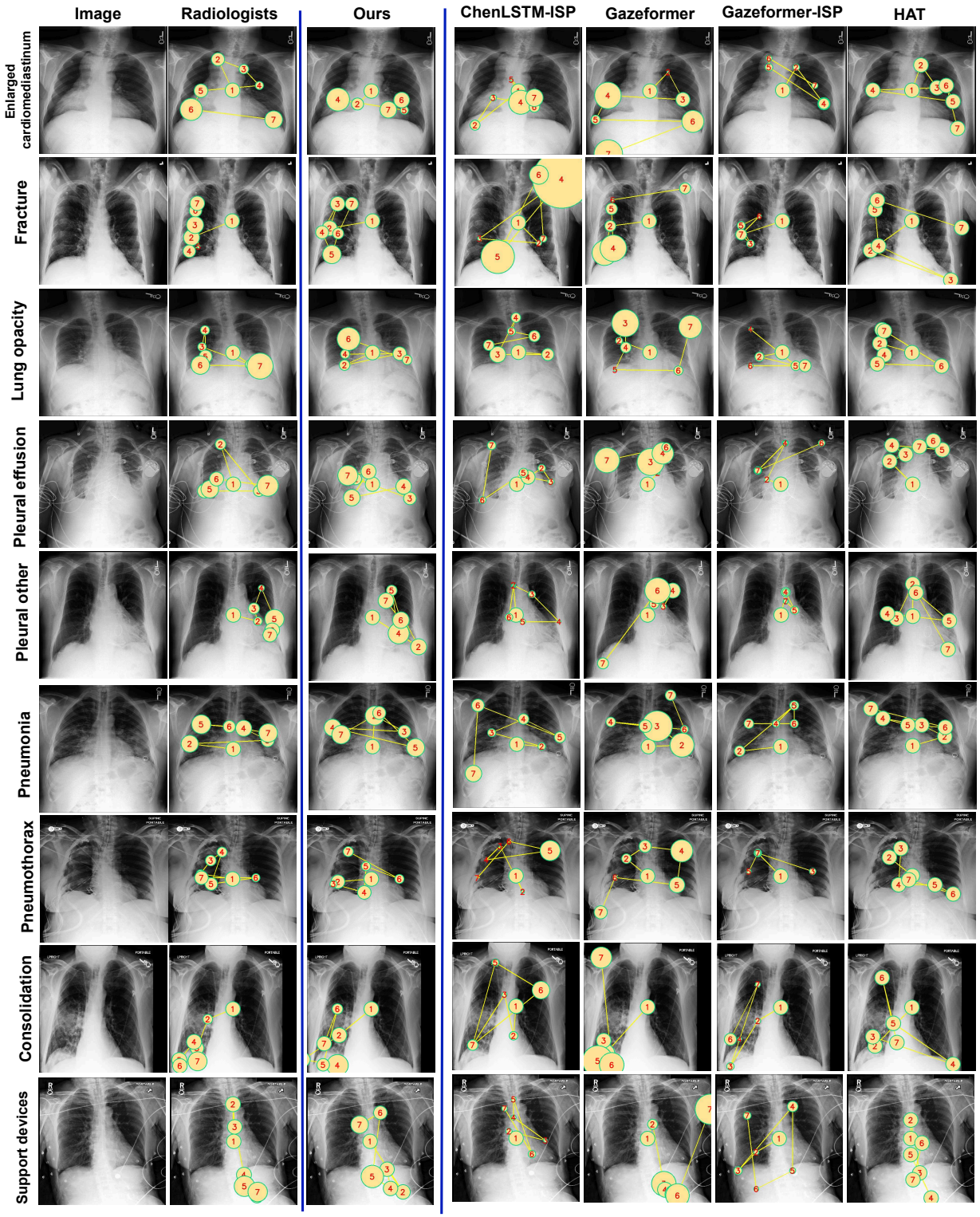



Figure 3. Additional qualitative results of ChestSearch (Ours), Gazeformer, GazeformerISP, ChenLSTMISP, and HAT. Each circle represents a fixation, with the number and radius indicating its order and duration, respectively. As HAT only predicts 2D coordinates, we let all predicted fixations of HAT have the same radius.

- Third row (Consolidation): "There is persistent airspace consolidation in the right mid lung and right medial lung base which may reflect residual pneumonia/aspiration."
- Fourth row (Edema): "There are indistinct pulmonary vascular markings suggestive of a component of interstitial edema. Findings suggestive of mild interstitial edema."
- Fifth row (Lung lesion): "Second nodular opacity in the left upper lung field is unchanged from ___. A nodular opacity in the left upper lung better delineated on the CT scan from ___ is concerning for malignancy and should undergo further diagnostic workup."
- Sixth row (Pneumothorax): "Right pneumothorax post surgery with three chest tubes in place."

Figure 5 compares ChestSearch with ChenLSTM-ISP, Gazeformer, Gazeformer-ISP, and HAT on other samples. Here are the radiology reports:

- First row (Lung lesion): "Persistent ill-defined nodular opacities with an upper lobe predominance, which may be slightly progressed when compared to the prior study."
- Second row (Lung lesion): "Stability of the right middle lobe calcified nodule."
- Third row (Pneumothorax): 'There is a 8-10 mm right apical pneumothorax without evidence of tension."
- Fourth row (Pneumothorax): "There has been no significant interval change in a small left apical pneumothorax."
- Fifth row (Support devices): "Left-sided pacer is noted with leads terminating in the right atrium and right ventricle."
- Sixth row (Atelectasis): "Streaky bibasilar opacities likely reflect atelectasis."
- Seventh row (Atelectasis): "Bibasilar atelectasis is mild to moderate."
- Eighth row (Atelectasis): "Dense consolidation that developed in the left lower lobe on ___ is unchanged, either atelectasis or pneumonia, but likely related to aspiration. There is more atelectasis at the right lung base today and early interstitial edema at both lung bases has worsened."
- Nineth row (Cardiomegaly): "The cardiac silhouette is mildly enlarged."

Figure 6 also compares ChestSearch with ChenLSTM-ISP, Gazeformer, Gazeformer-ISP, and HAT on other samples. Here are the radiology reports:

- First row (Cardiomegaly): "Continued enlargement of the cardiac silhouette with mild elevation of pulmonary venous pressure and bibasilar atelectatic changes."
- Second row (Cardiomegaly): "The heart size is enlarged."
- Third row (Consolidation): "There is increased retrocardiac density consistent with left lower lobe collapse and/or consolidation, unchanged."
- Fourth row (Consolidation): "No change in the right lower lung consolidation is demonstrated. Left basal opacity appears to be unchanged as well, less extensive as compared to the right lower lobe."
- Fifth row (Consolidation): "Right lower lobe pneumonic consolidation is unchanged."
- Sixth row (Fracture): "Multiple left-sided rib fractures are again noted."
- Seventh row (Lung lesion): "Right hilar opacity likely related to pneumonia though underlying lesion cannot be excluded."
- Eighth row (Lung lesion): "Left upper lobe collapse is due to a large lobulated left hilar mass obstructing the upper lobe bronchus."
- Nineth row (Pneumothorax): "Tiny right apical pneumothorax is newly apparent."

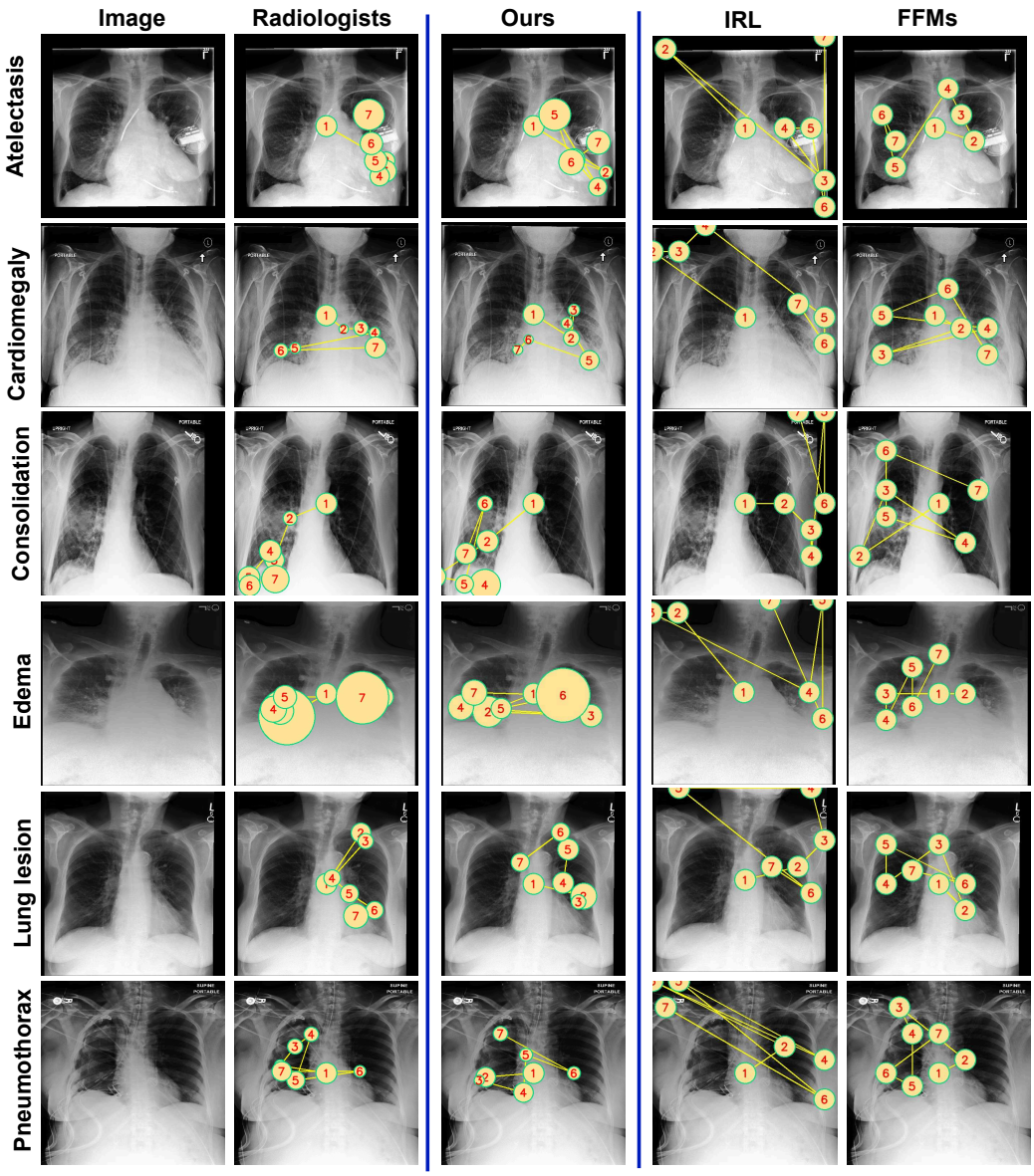



Figure 4. Additional qualitative results of ChestSearch (Ours), IRL, and FFMs. IRL and FFMs only predict 2D coordinates, the same as HAT, so we let their fixations have the same radius.

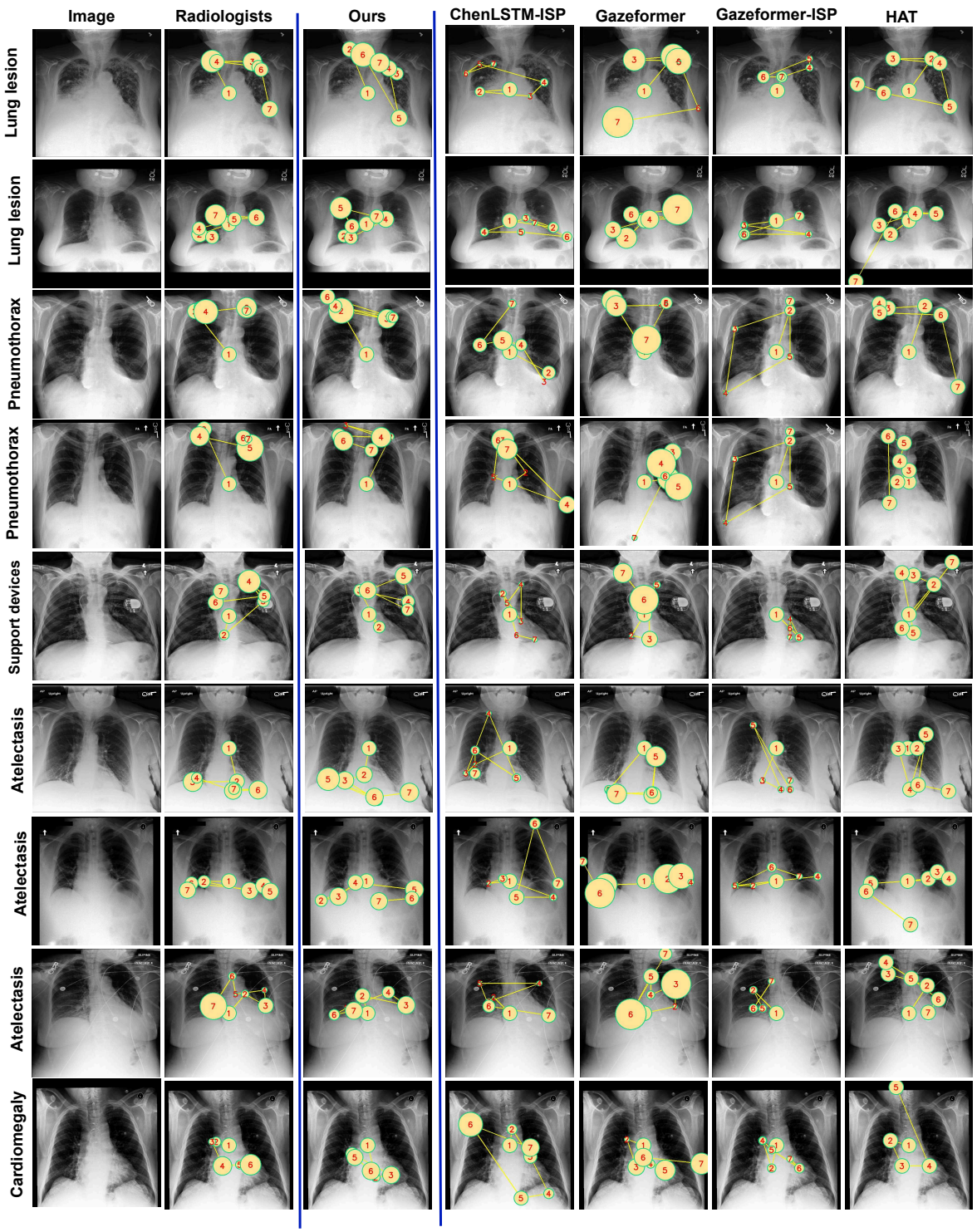



Figure 5. Additional qualitative results of ChestSearch (Ours), Gazeformer, GazeformerISP, ChenLSTMISP, and HAT. Findings are Cardiomegaly, Atelectasis, Support Devices, Lung lesion, and Pneumothorax.

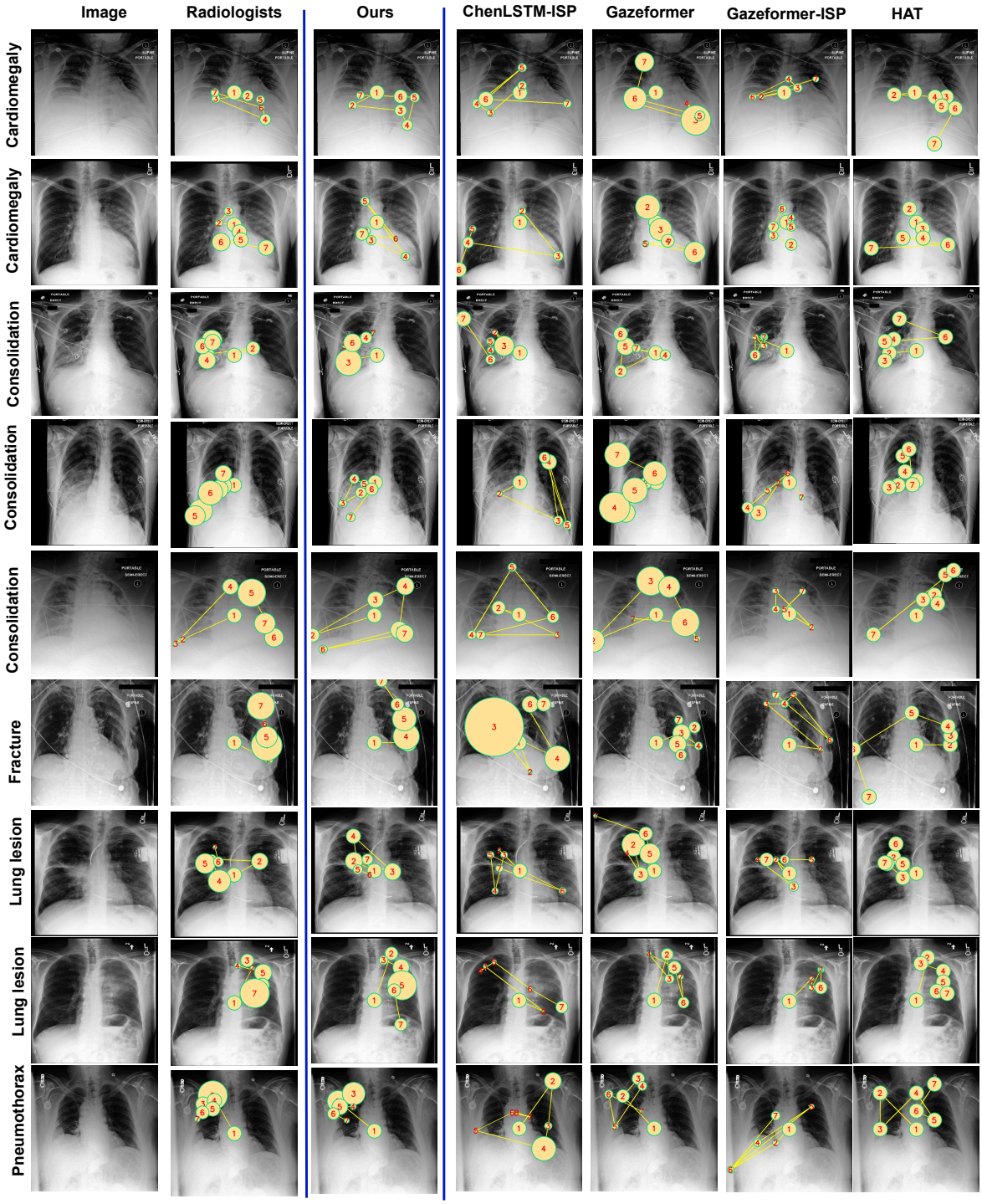



Figure 6. Additional qualitative results of ChestSearch (Ours), Gazeformer, GazeformerISP, ChenLSTMISP, and HAT on another set of samples. Findings are Cardiomegaly, Consolidation, Fracture, Lung lesion, and Pneumothorax.

## References


[1] 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. 1

[2] 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. 3

[3] 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. 3

[4] Andrew T Duchowski and Andrew T Duchowski. *Eye tracking methodology: Theory and practice*. Springer, 2017. 1

[5] 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. 3

[6] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019. 2

[7] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, 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. 1, 2

[8] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[9] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, 45(1), 2013. 2

[10] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[11] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. 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. 3

[12] H Nguyen, HH Pham, NT Nguyen, DB Nguyen, M Dao, V Vu, K Lam, and LT Le. Vinbigdata chest x-ray abnormalities detection. *Kaggle Competition https://www. kaggle. com/c/vinbi gdatachest-xray-abnor malit ies-detec tion*, 2020. 3

[13] 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

[14] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 2

[16] 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

[17] 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. 3

[18] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 3