CT-ScanGaze: A Dataset and Baselines for 3D Volumetric Scanpath Modeling

Trong Thang Pham[1], Akash Awasthi[2], Saba Khan[2], Esteban Duran Marti[1], Tien-Phat Nguyen[3], Khoa Vo[1], Minh Tran[1], Ngoc Son Nguyen[4], Cuong Tran Van[4], Yuki Ikebe[1], Anh Totti Nguyen[5], Anh Nguyen[6], Zhigang Deng[2], Carol C. Wu[7], Hien Nguyen[2], and Ngan Le[1]

[1]University of Arkansas, [2]University of Houston, [3]University of Science VNU-HCM, [4]FPT Software, [5]Auburn University, [6]University of Liverpool, [7]MD Anderson Cancer Center

Abstract

Understanding radiologists' eye movement during Computed Tomography (CT) reading is crucial for developing effective interpretable computer-aided diagnosis systems. However, CT research in this area has been limited by the lack of publicly available eye-tracking datasets and the three-dimensional complexity of CT volumes. To address these challenges, we present the first publicly available eye gaze dataset on CT, called CT-ScanGaze, captured from expert radiologists. Then, we introduce CT-Searcher, a novel 3D scanpath predictor designed specifically to process CT volumes and generate radiologist-like 3D fixation sequences, overcoming the limitations of current scanpath predictors that only handle 2D inputs. Since deep learning models benefit from a pretraining step, we develop a pipeline that converts existing 2D gaze datasets into 3D gaze data to pretrain CT-Searcher. Through both qualitative and quantitative evaluations on CT-ScanGaze, we demonstrate the effectiveness of our approach and provide a comprehensive assessment framework for 3D scanpath prediction in medical imaging. Code and data are available at https://github.com/UARK-AICV/CTScanGaze.

1. Introduction

Interpretability is a fundamental aspect of Computer-aided Diagnosis (CAD) system as it supports safe, accurate, and trustworthy patient care [47]. The integration of eye gaze signals offers a promising approach to enhance the interpretability of CAD systems [22, 38, 47, 48, 51]. Understanding the importance of gaze data in medical imaging analysis, EGD [38] and REFLACX [6] are created and shared publicly to advance chest X-rays (CXR) analysis. As shown in Fig. 1a, these datasets have facilitated developments for artificial intelligent solutions, including multi-label classification [38, 61, 68], report generation [38, 60], explainable classification [55], explainable report generation [54], and scanpath prediction [66]. However, existing medical datasets are either limited to 2D images like CXRs or lack radiologist gaze data. The absence of public dataset capturing radiologists' eye gaze patterns on Computed Tomography (CT) has left several critical tasks underexplored in CT imaging, including scanpath prediction, explainable

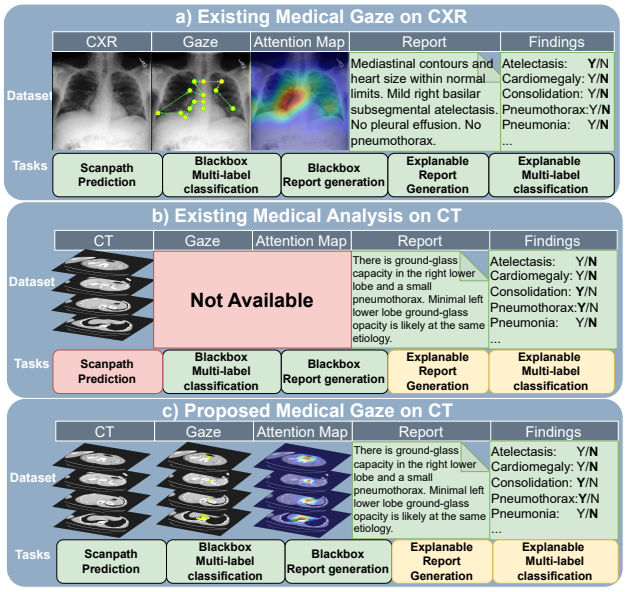

Figure 1. Many research directions in CAD would benefit from the availability of gaze, report, and findings data, as is the case for CXR (a), highlighted in green. However, some critical areas in CT research are underexplored. For example, there are only preliminary results for Explainable Report Generation and Explainable Classification, highlighted in yellow. Especially, Scanpath Prediction is underexplored and often overlooked, highlighted in red, primarily due to the lack of publicly available datasets (b). Our dataset offers new research opportunities to these tasks. In this paper, we address the Scanpath Prediction task on CT scans (c).

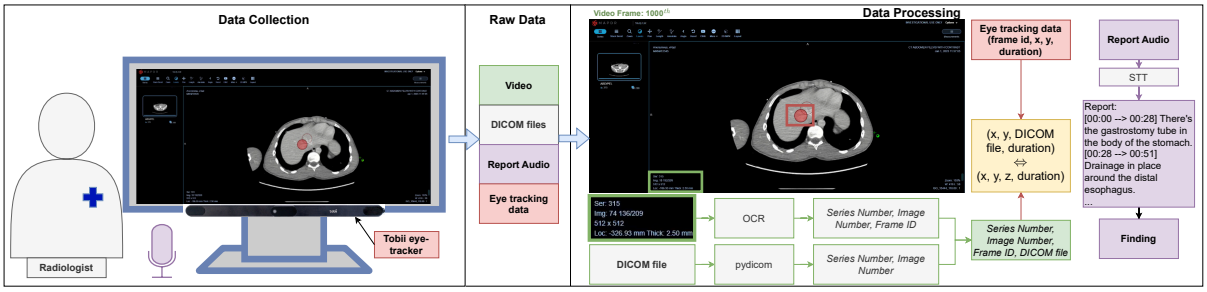

Figure 2. Illustration of our data collection and processing pipeline. The Data Collection panel shows the setup: a radiologist examines CT scans on a monitor equipped with a Tobii eye-tracker and microphone for recording audio reports. The Raw Data panel displays four data streams collected: screen recording video, DICOM files, verbal report audio, and eye gaze data. The Data Processing panel demonstrates the data processing pipeline, which includes OCR and pydicom processing of DICOM files, integration of eye gaze data (frame ID, (x, y) coordinates, duration) to create the final 3D gaze data. We create the final radiology report and clinical findings from the Report Audio.

diagnosis (classification and report generation), as shown in Fig. 1b. Especially, the scanpath prediction task on CT data is often overlooked due to missing gaze data.

The volumetric nature of CT data reveals several unique behaviors from radiologists' eye movement. For example, radiologists must constantly navigate across multiple slices and viewing planes to understand anatomical relationships [2, 22]. Additionally, radiologists mentally integrate spatial information across different depths to form a cohesive 3D understanding of the anatomy and pathology [2, 21, 24]. Finally, they employ systematic search strategies to ensure thorough volume examination, as overlooking even a single slice risks missing critical findings [2, 22, 67]. The existing scanpath prediction methods [13, 50, 79] are primarily designed for 2D imaging analysis. Consequently, these 2D-based models do not account for the unique aspects of CT interpretation, particularly the complex volumetric eye movement strategies and comprehensive slice coverage patterns exhibited by radiologists, such as moving back-and-forth between slices. Motivated by these challenges, this paper introduces the first public eye gaze medical dataset that focuses on CT scans, CT-ScanGaze. Unlike existing medical eye gaze datasets [6, 38], CT-ScanGaze provides 3D eye gaze data associated with every CT volume. As shown in Fig. 1c, CT-ScanGaze provides four main modalities: CT scans, eye gaze data (gaze map and attention map), radiology reports, and findings.

To conduct a benchmark on the proposed CT-ScanGaze, we tackle the scanpath prediction task. While existing scanpath prediction methods can be extended to 3D, their original 2D design may limit their ability to model inter-slice navigation and spatial-temporal continuity. Moreover, scaling these methods to 3D introduces a higher-dimensional search space, making them more susceptible to the curse of dimensionality and harder to generalize without specialized architectural design. Therefore, we propose a transformer-based network, CT-Searcher, that generates 3D scanpaths for CT volumes. DL typically requires large pretraining datasets, but CT-ScanGaze is relatively small with only 909 CT volumes compared to COCO-Search18's 6,202 images, potentially leading to overfitting or suboptimal training. So, we utilize a 2D-to-3D pipeline to create a synthetic 3D gaze dataset for the pretraining step. By pretraining CT-Searcher, it gains the ability to process CT features and predict gaze-like sequences, which enhances its performance on the final scanpath prediction task. Next, we train and evaluate CT-Searcher against current state-of-the-art scanpath prediction models [13, 50, 79], which are adapted to process 3D inputs and predict 3D scanpaths. Our findings indicate that CT-Searcher successfully generates radiologist-like scanpaths, effectively capturing both spatial coverage within each CT slice and navigational movements across slices.

Our contributions are summarized as follows:

- **CT-ScanGaze:** We present the first public dataset of expert radiologist gaze during CT analysis, comprising CT scans, eye gaze data, detailed reports, and findings.
- **CT-Searcher:** We introduce a 3D scanpath prediction network that generates radiologist-like eye movement from a CT volume. And a pretraining pipeline on synthetic 3D gaze data (CT, 3D gaze) from 2D gaze data (CXR, 2D gaze).

2. CT-ScanGaze

2.1. Data Collection

Our study collects data from a private hospital dataset of chest and abdomen CT scans, working with two experienced radiologists (10+ years experience). Data collection setup involves a Tobii eye tracker mounted on a monitor to track gaze data and a microphone to record audio reports. Calibration is performed before each session, with Tobii Pro Lab software managing the eye-tracking data and screen recordings. CT scans are viewed using OHIF viewer integrated with OHIF-Orthanc server, providing necessary

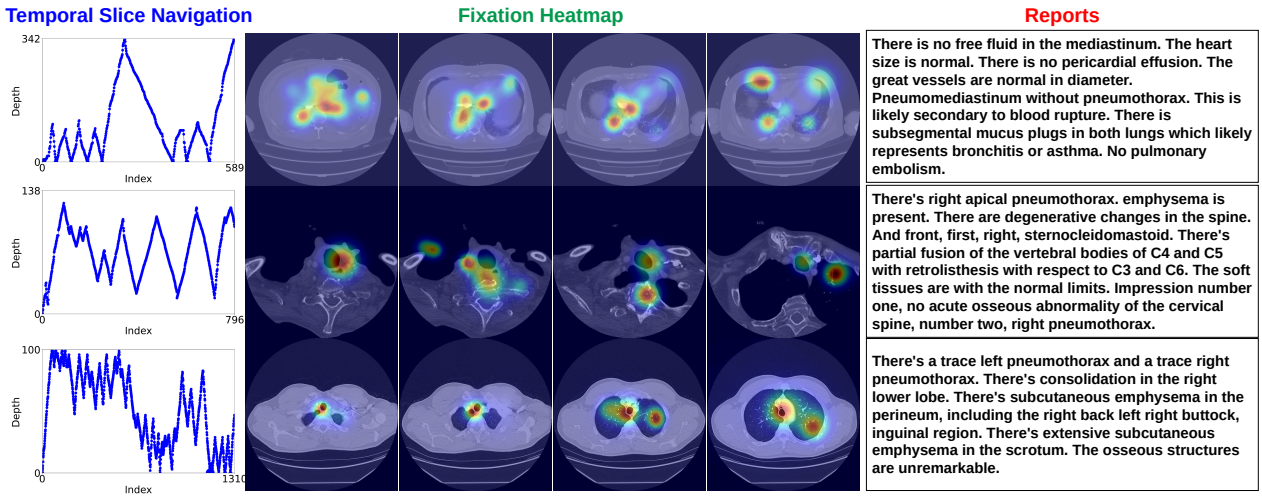

Figure 3. Examples from our dataset. Three CTs are reviewed and concluded with radiology reports as shown in the **Reports** column on the right. The **Temporal Slice Navigation** shows how slice navigation change over time. We can observe that radiologists often scan through all slices and go back-and-forth for suspicious areas. In the middle column **Fixation Heatmap**, we show some slices in the CTs. The heatmap represents all fixations viewed on that slice. For example, if in the first view they only see the left side, and in the second view they see the right side, the visualized heatmap shows fixations on both sides.

tools, e.g. contrast window control, for radiological analysis. We organize the setup as in Fig. 2 (left). Our radiologists perform their natural reading process of every CT scan exactly like they do in clinical practice. For example, 5mm images are sufficient for many pathologies. It is impractical for radiologists to review too many images in the thin series. Thin series are used for troubleshooting when radiologists need to examine certain details. Therefore, our CT scans vary in slice thickness (1-5 mm) to match this practice.

2.2. Data Processing

Each data collection session yields four data streams: video recording of radiologists' CT reading session, DICOM files of the CT scan, audio recording report, and eye gaze data. As shown in Fig. 2 (right), we process these streams to create through three main steps: Spatial Mapping, Spatiotemporal Mapping, Radiological Report & Findings Extraction. Fig. 3 shows 3 examples being viewed by our radiologists.

Spatial Mapping. We extract individual frames from session videos and use Tesseract OCR [63] to identify series and slice numbers from the bottom-left text, with manual correction when needed. This creates frame-to-DICOM mapping pairs (Fig. 2).

Spatiotemporal Mapping. The eye tracker records time (ms), screen coordinates (x, y), and fixation duration (t). Each timestamp is mapped to a frame ID using the video's frame rate (25 FPS). Using frame-to-DICOM pairs, we map gaze data to DICOM files to obtain (DICOM ID, (x, y, t)) pairs. Since each DICOM represents a CT slice number, we transform fixations into a list of 4-tuples (x, y, z, t), where z is a slice number.

Radiological Report & Findings Extraction. We use Google's Speech-to-Text 'medical dictation' model [14] on the recorded audio to generate a textual report of the radiologist' verbal interpretation for each CT scan. From these reports, we use SARLE [20], a specialized labeler for extracting findings from CT reports, to extract the corresponding radiological findings.

2.3. Dataset Statistics

CT-ScanGaze contains 909 CT scans, each accompanied by: scanpath, a radiology report, and findings. We have a total of 131,618 CT slices, 4,772 minutes of scanpath data, and 9,332 extracted findings. Due to the original gaze data containing dense and complex scanpaths with sequences averaging 543 fixations (and reaching up to 2,708 fixations per CT), we employ a simplification algorithm from the MultiMatch toolbox [18] to make the data more manageable while preserving essential gaze patterns. This process reduces sequences to an average of 222 fixations with a maximum of 1,507 fixations. Both original and simplified versions will be made available. We recommend the reader to see Appendix A for more statistical details and Appendix H for a discussion on the gaze simplification algorithm.

2.4. Scientific Benefits

This dataset represents a valuable resource for the medical imaging and computer vision communities:

- **Benchmark for 3D Scanpath Prediction:** CT-ScanGaze serves as a benchmark for predicting visual search patterns in 3D medical volumes, addressing difficulties

unique to volumetric imaging that existing 2D-focused datasets do not cover.

- **Advancement of Explainable AI**: By linking radiologists' gaze with radiology diagnosis, CT-ScanGaze supports research in explainable report generation and classification, enhancing insights into the connection between visual attention and diagnostic reasoning. This research area is actively explored in 2D medical imaging analysis by various researchers [6, 32, 38, 57, 70].

In addition, we believe CT-ScanGaze can be used in other scenarios such as study of radiologists' gaze behavior [22], advancement of general 3D scanpath prediction, radiology training [32, 38, 70], and collaborative CAD [39, 51, 53].

3. Problem Statement: 3D Scanpath Prediction

In this work, we address the 3D scanpath prediction task on CT scans. Given a CT volume $V \in \mathbb{R}^{H\times W\times D}$, our goal is to predict a sequence of N fixations $\{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_N\}$. Each predicted fixation $\hat{p}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i)$ should minimize its deviation from the ground truth fixation sequence $\{p_1, p_2, \ldots, p_N\}$, where each $p_i = (x_i, y_i, z_i, t_i)$ represents the 3D spatial location (x_i, y_i, z_i) and duration t_i of a radiologist's gaze point. Our model aims to capture both the spatial patterns and temporal dynamics of expert visual search behavior in volumetric medical imaging.

4. CT-Searcher

The overall architecture of CT-Searcher is depicted in Fig. 4. First, a Feature Extraction module takes a CT volume V and produces suitable 3D-aware representations (Sec. 4.1). Then, a Transformer Decoder uses a set of learnable queries to attend on the 3D-aware presentation and creates appropriate decoded features for decoding our desired outputs (Sec. 4.2). Finally, the latent features go through a Spatial Prediction module to produce 3D probability maps (Sec. 4.3) and a Duration Prediction module to produce durations (Sec. 4.4). All modules in CT-Searcher are trained jointly with losses defined in Sec. 4.5.

4.1. Feature Extraction

Given an input CT volume $V \in \mathbb{R}^{H\times W\times D}$, we first extract visual features $F \in \mathbb{R}^{H'\times W'\times D'\times C}$ using a visual encoder, where H', W', and D' are the reduced spatial dimensions and C is the feature dimension. These features are then projected into a compatible representation $Z = \text{MLP}(cc(F, \tau)) \in \mathbb{R}^{L\times D_m}$ for the Transformer Encoder by a Multi-Layer Perceptron (MLP), where $cc(\cdot,\cdot)$ is a concatenate operation, τ is a special learnable token representing 'stop' fixation, D_m is the transformer hidden dimension and $L = H' * W' * D' + 1$. To incorporate spatial information, we adapt the 2D sinusoidal positional encoding [8] to a 3D version. For each dimension (x, y, z) in a 3D volume, we generate position embeddings using different frequency bands:

$$\omega_k = \exp\left(-k\frac{\log(T)}{d/2}\right), \quad k = 0, \ldots, d/2 - 1 \tag{1}$$

where $T = 10000$ is the temperature parameter and $d = D_m/3$ is the embedding dimension per axis. For each spatial position pos along each axis, we compute:

$$PE_{axis}(pos) = (\sin(pos \cdot \omega_0), \cos(pos \cdot \omega_0), \ldots, \sin(pos \cdot \omega_{N-1}), \cos(pos \cdot \omega_{N-1})) \tag{2}$$

where N is the maximum length of fixation, $axis \in \{x, y, z\}$ and positions pos are normalized to $[0, 2\pi]$. The final encoding $PE(x, y, z)$ concatenates these components:

$$PE(x, y, z) = [PE_x(x); PE_y(y); PE_z(z)] \in \mathbb{R}^{D_m} \tag{3}$$

We apply 3D positional encoding on only the spatial tokens $H' * W' * D'$ of Z and pass Z to a Transformer Encoder that composes suitable 3D-aware representations $E(Z) \in \mathbb{R}^{L\times D_m}$ for the Transformer Decoder in the next step.

4.2. Transformer Decoder

Before going through Transformer Decoder, we apply 3D positional encoding again on $E(Z)$ to retain the 3D spatial information. Then, a Transformer Decoder $D(\cdot)$ uses N learnable gaze queries $Q \in \mathbb{R}^{N\times D_m}$ to attend to relevant features in $E(Z)$ to produce decoded features $R = D(Q, E(Z)) \in \mathbb{R}^{N\times D_m}$.

4.3. Spatial Prediction

The objective of Spatial Prediction module (SP) is to generate 3D fixation maps that represent the probabilistic distribution of 3D fixation coordinates over time and the the likelihood of sequence termination. The output of the transformer's decoder $R \in \mathbb{R}^{N\times D_m}$ is projected into a fixation embedding by a Fully-connected (FC) layer, which is then convolved with the encoded feature map $E(Z)$ and a softmax layer to get the 3D spatial distribution for all fixations $\hat{Y} \in [0, 1]^{N\times L}$:

$$\hat{Y} = \text{softmax}(\text{FC}(R) \otimes E(Z)^\top), \tag{4}$$

where $\otimes$ denotes the matrix multiplication.

4.4. Duration Prediction

The objective of Duration Prediction module (DP) is to generate fixation durations that reflect the probabilistic distribution of fixation durations over time. Here, we use the re-parameterization trick [19] on the encoded features R to regress them into mean values μ_t and log-variances λ_t:

$$\mu_t = \text{MLP}_{\mu_t}(R), \quad \lambda_t = \text{MLP}_{\lambda_t}(R) \tag{5}$$

$$\hat{t} = \mu_t + \epsilon_t \cdot \exp(0.5\lambda_t), \tag{6}$$

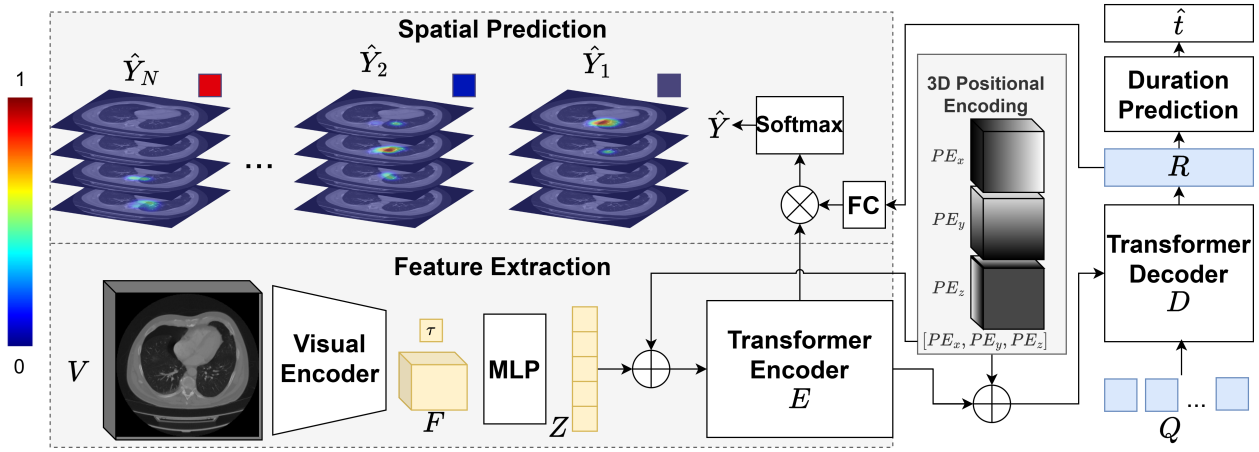

Figure 4. CT-Searcher processes CT scans V to predict 3D scanpaths. Initially, a 3D visual encoder within the Feature Extraction module extracts voxel features F. These features with a special 'stop' token are then transformed into Z by an MLP and combined with 3D positional encoding PE to incorporate spatial information. The Transformer Encoder subsequently composes suitable 3D-aware representations from Z. A Transformer Decoder then uses a set of learnable queries Q to attend to these 3D-aware representations and generate decoded features R for the desired outputs. To prevent the loss of 3D spatial information, we reapply the 3D positional encoding before feeding the 3D-aware representations into the Transformer Decoder. The decoded features R are then processed by the Spatial Prediction to produce $\hat{Y}$, including 3D fixation maps and a 'stop' probability. R is also processed by the Duration Prediction Head to estimate fixation durations $\hat{t}$ over time. Each $\hat{Y}$ has $H * W * D + 1$ elements, where $H * W * D$ tokens represent the 3D fixation map and one special element, embedded from token τ, represents 'stop'. In the final $\hat{Y}_N$ of this figure, the special token has the highest probability and is likely to be chosen in sampling, signaling that the model predicts fixation stops at the N^{th} fixation.

where $\epsilon_t \sim \mathcal{N}(0, 1)$ is a noise term that give our predictions a probabilistic characteristic.

4.5. Losses

To model the probabilistic nature of human fixations, we use two complementary loss functions. For the SP, which generates 3D fixation probability maps, we use Cross Entropy loss between the predicted distribution $\hat{Y}_i$ and ground truth fixation map $Y_i \in [0, 1]^{L_{gt}}$ at each step i:

$$\mathcal{L}_{ce} = \frac{-1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L_{gt}} Y_i(l) \log \hat{Y}_i(l) \tag{7}$$

where $L_{gt} = H * W * D + 1$ are the resolution of the ground truth volumes. $\hat{Y}_i$ is interpolated to same shape $Y_i \in \mathbb{R}^{H*W*D+1}$ before computing loss. The i-th ground truth 3D map U_i is initialized as zero, with the fixation location (x, y, z) as 1:

$$U_i(x', y', z') = \begin{cases} 1 & \text{if } x' = x, y' = y, z' = z \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $x' \leq W, y' \leq H, z' \leq D$. We use 1-hot ground-truth Y framing as classification task, and CTSearch predicts distribution $\hat{Y}$ with softmax in Eq. (4). Because N is fixed, we use padding on sequences with length less than N and set the stop token τ to be 1 and $U_i = 0$, otherwise $\tau = 0$. The ground truth heatmap is $Y_i = cc(U_i, \tau)$.

The output of DP module is a scalar, making it suitable for a regression objective. We observe that the L_1 loss performs effectively in our scenario.

$$\mathcal{L}_t = \frac{1}{N} \sum_{i=1}^{N_p} \|\hat{t}_i - t_i\|_1 \tag{9}$$

where t_i is the ground truth duration at step i. The final loss combines both terms:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_{ce} \tag{10}$$

4.6. Pre-training CT-Searcher

Due to the high complexity of 3D scanpath prediction, we find that pretraining CT-Searcher before training on CT-ScanGaze is necessary to enhance its ability to process CT features and predict fixation-like sequences.

First, we combine the eye gaze data from EGD [38] and REFLACX [6], which brings more than 3,000 pairs of CXR and fixations. We also remove invalid fixation data described by the authors [6, 38] of EGD and REFLACX.

Secondly, we use a CXR-to-CT method [40] to convert all CXRs to CTs. To transform 2D gaze points into 3D coordinates, we process each normalized 2D fixation set $\{(x_i, y_i, t_i)\}_{i=1}^{n}$, where $x_i, y_i \in [0, 1]$ represent normalized coordinates in the CXR image plane (with $(0, 0)$ at the top-left corner and $(1, 1)$ at the bottom-right corner), and t_i is the fixation duration. We perform the following steps: flip

the x dimension using $1 - x_i$ to account for the right-to-left reading pattern in CXR; set the middle slice position as 0.5 (where 0 represents the first slice and 1 represents the last slice in normalized coordinates); and map y_i directly to the slice dimension while preserving the normalized scale. The middle slice position (0.5) is chosen because radiologists exhibit a center bias when reading axial slices. This process transforms the 2D fixation set $\{(x_i, y_i, t_i)\}_{i=1}^{n}$ into a 3D fixation set $\{(1-x_i, 0.5, y_i, t_i)\}_{i=1}^{n}$ with approximately 3,000 samples.

Finally, we augment the data during pretraining by sampling fixations from a Gaussian distribution with σ of one degree of visual angle, which fluctuates coordinates in the first and second dimension. This follows the widely accepted assumption [43] that fixations follow a Gaussian distribution. We emphasize that this conversion process should only be used during pretraining due to the inherent domain gap between 2D and 3D imaging. We provide additional synthetic visualization in Appendix B.

5. Experiments

5.1. Experimental Details

Implementation Details. CT-Searcher uses Swin UNETR [29] encoder with $96 \times 96 \times 96$ input window. Due to computational constraints, we use a frozen pre-trained Swin UNETR checkpoint on LIDC-IDRI [3]. The transformer has 6 encoder/decoder layers, 8 attention heads, and hidden size of 768. All MLPs in CT-Searcher have two layers, ReLU activation, and 512 hidden dimension. The CT-Searcher is implemented in PyTorch. We use AdamW [46] optimizer with initial learning rate of $1e^{-4}$ and weight decay of 0.01. We train and evaluate all methods on the simplified version (Sec. 2.3) of CT-ScanGaze. We pre-train CT-Searcher (Sec. 4.6) for 50,000 iterations with a batch size of 8, and we then fine-tune CT-Searcher on CT-ScanGaze for 20 epochs with $N = 400$ fixations. For more implementation details, please refer to Appendix C.

Evaluation Metrics. We benchmark on two aspects: **scanpath-based** metrics (SM [16], MM [18], SED [7, 27]) and **spatial-based** metrics (CC, KLDiv, NSS [13, 30]). All metrics are adapted for 3D scanpath prediction (Appendix D). We run 5-fold cross-validation and report the scores with 95% confidence intervals.

Baselines. We evaluate CT-Searcher against leading scanpath prediction methods including PathGAN [4] Gazeformer [50], GazeformerISP [13] and HAT [79]. To benchmark these 2D methods on 3D gaze data, we replace the feature encoder of PathGAN, Gazeformer, GazeformerISP, and HAT to Swin UNETR encoder. We replace 2D positional encoding in Gazeformer, GazeformerISP, and HAT to our 3D positional encoding. Because Gazeformer has two heads for predicting the 2D coordinates and one head for predicting the duration, we only add a head to predict the slice number. In HAT and GazeformerISP, we alter their spatial decoder from generating 2D fixation heatmaps to 3D fixation heatmaps. For remaining modules, we follow their original implementation, for example we use RoBERTa [45] for task embedding of Gazeformer and GazeformerISP. More implementation details on 3D adapted scanpath prediction baselines are in Appendix E.

5.2. Qualitative Results

We present qualitative results in Fig. 5 to demonstrate the effectiveness of our proposed method. The results reveal clear performance differences across evaluation aspects. In the temporal slice navigation plots (Fig. 5, left), CT-Searcher captures similar temporal dynamics to the ground-truth, while baseline methods (GazeformerISP and HAT) exhibit erratic, high-variance patterns. Notably, the **orange areas** show that CT-Searcher replicates the non-linear behavior characteristic of experienced readers, who frequently navigate back and forth through the volume to revisit suspicious regions. The fixation heatmap visualization (Fig. 5, right) demonstrates that CT-Searcher produces more accurate attention distributions compared to other methods. These qualitative findings confirm that our approach successfully replicates the complex visual search patterns of radiologists during CT interpretation, representing a significant advancement in 3D medical scanpath prediction. We provide additional qualitative results in Appendix F.

5.3. Quantitative Results

Tabs. 1 and 2 demonstrates the significant challenges in 3D scanpath prediction. CT-Searcher achieves consistently better performance across all metrics compared to existing approaches. The earliest method, PathGAN, performs poorly across all metrics, with particularly low ScanMatch (0.0118) and saliency scores (CC: 0.0349). Recent transformer-based approaches show improvement but face various limitations: HAT struggles with limited training data (ScanMatch: 0.0171), Gazeformer's separate coordinate prediction via three MLPs causes consistency issues affecting saliency metrics (KLDiv: 23.332), and GazeformerISP underperforms on CT-ScanGaze due to its specialized and complex architecture. CT-Searcher demonstrates superior performance with significant improvements in ScanMatch (0.1466), MultiMatch position (0.7859) and duration (0.5003) scores, and saliency metrics (CC: 0.1706, KLDiv: 3.645), indicating better capture of radiologists' gaze patterns. These results highlight the effectiveness of our approach and represent a substantial advancement in modeling expert visual search behavior for CT.

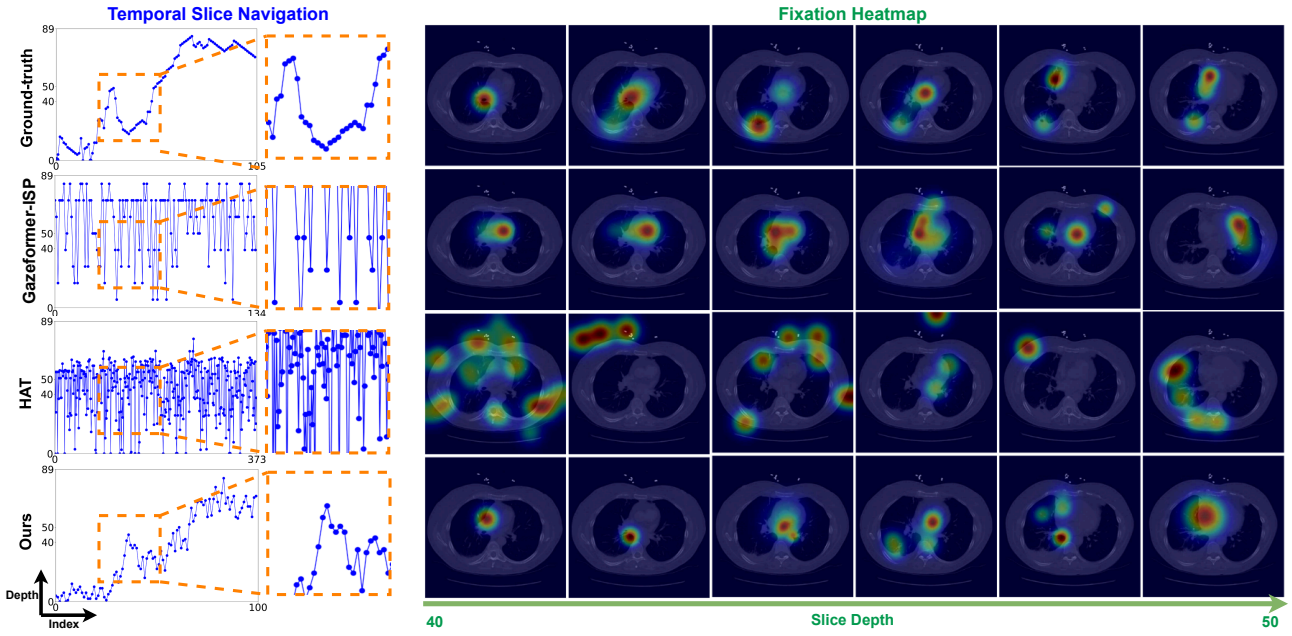

Figure 5. Qualitative comparison of our method and other SOTA methods, HAT and GazeformerISP. The **Temporal Slice Navigation** column shows the predicted scanpath over time, with the y-axis representing slice numbers (out of 89 total) and the x-axis showing fixation indices. For example, the ground truth has 105 fixations, with fixations move through depth from the earliest to the last slices. The **Fixation Heatmap** columns present the fixation heatmaps, illustrating regions of interest from which the corresponding gaze positions are sampled. The visualized CT slices for these heatmaps are selected between slice numbers 40 and 50. We observe that CT-Searcher can capture the navigation pattern of radiologists. Especially, the **orange area** indicates that the back-and-forth scanning behavior of radiologists has been successfully captured by CT-Searcher.

Table 1. Comparison of scanpath-based metrics between our CT-Searcher and existing models adapted to 3D data. (Appendix E).

Method	ScanMatch ↑		MultiMatch ↑					SED ↓
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	
PathGAN [4]	0.0118±0.002	0.0649±0.005	0.8277±0.023	0.3194±0.015	0.6786±0.031	0.6559±0.028	0.2959±0.018	663±42
HAT [79]	0.0171±0.003	-	0.8103±0.019	0.3178±0.014	0.6522±0.029	0.6295±0.025	-	307±28
Gazeformer [50]	0.0619±0.005	0.0718±0.006	0.8653±0.021	0.3012±0.016	0.8601±0.035	0.6492±0.027	0.3254±0.019	279±25
GazeformerISP [13]	0.0828±0.007	0.0711±0.006	0.8831±0.024	0.3060±0.015	0.8044±0.033	0.7354±0.031	0.3375±0.020	238±21
Ours CT-Searcher	**0.1466±0.009**	**0.1170±0.008**	**0.9216±0.026**	**0.4151±0.018**	**0.8783±0.036**	**0.7859±0.033**	**0.5003±0.024**	**174±18**

5.4. Cross-radiologist Evaluation

To assess potential learning bias toward a single radiologist's style (as CT-ScanGaze is collected from only two radiologists), we conduct a cross-radiologist evaluation. Using the same trained checkpoints from our 5-fold cross validation that produce the results in Tabs. 1 and 2, we compute separate scores for test sets containing only the first or second radiologist's ground truth data, as shown in Tabs. 3 and 4. The differences in scores between the two radiologists are insignificant, leading us to conclude that CT-Searcher successfully learns general scanpath patterns rather than overfitting to an individual radiologist's style.

Table 2. Comparison of spatial-based metrics between different scanpath prediction models.

Method	Saliency		
	CC ↑	KLDiv ↓	NSS ↑
PathGAN [4]	0.0349±0.017	25.602±1.832	0.1343±0.021
HAT [79]	0.0914±0.034	15.493±1.245	1.0676±0.089
Gazeformer [50]	0.0855±0.044	23.332±1.756	0.5005±0.042
GazeformerISP [13]	0.1104±0.021	5.023±0.428	0.7703±0.065
Ours CT-Searcher	**0.1706±0.016**	**3.645±0.312**	**1.1422±0.095**

5.5. Ablation Studies

We conduct comprehensive ablation studies to validate each component of our framework. We provide additional ablation study on CT Visual Encoder backbone in Appendix G.

Impact of 3D Positional Encoding (3D-PE). 3D-PE enhances the model's ability to capture positional relationships across height, width, and depth. As shown in Tab. 5,

Table 3. Performance of our proposed CT-Searcher on different radiologists on scanpath-based metrics.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	
Radiologist #1	0.1499±0.010	0.1214±0.009	0.9087±0.025	0.4023±0.019	0.8888±0.035	0.8001±0.034	0.5110±0.023	182±19
Radiologist #2	0.1431±0.008	0.1124±0.007	0.9339±0.027	0.4273±0.017	0.8673±0.037	0.7712±0.032	0.4891±0.025	166±17
Both	0.1466±0.009	0.1170±0.008	0.9216±0.026	0.4151±0.018	0.8783±0.036	0.7859±0.033	0.5003±0.024	174±18

applying 3D-PE improves our model performance across all metrics. Comparing rows #1 and #2 reveals substantial improvements when adding 3D-PE without pretraining, while comparing rows #3 and #4 demonstrates that 3D-PE remains crucial even with pretraining in place. The dramatic improvement in KLDiv from 7.526 to 3.645 between rows #3 and #4 highlights how 3D-PE enhances spatial awareness in our volumetric predictions. In conclusion, 3D-PE provides essential spatial context for volumetric data processing, resulting in substantial performance improvements across all metrics regardless of whether pretraining is used.

Impact of Pre-training Step (Pre). Pretraining often benefits deep learning models by establishing helpful inductive biases [75]. Given the limited size of our real CT dataset, direct training alone may lead to suboptimal performance. As shown in Tab. 5, our experiments demonstrate that incorporating a pretraining step further improves model performance. When comparing rows #2 and #4, we observe that adding pretraining to a model with 3D-PE further enhances performance, with improvements in mSM (0.1275 to 0.1318), mMM (0.6883 to 0.7002), SED (184 to 174), and KLDiv (5.194 to 3.645). These gains appear across all evaluation metrics, confirming the value of pretraining.

6. Related Work

Eye Gaze Datasets. The proliferation of eye gaze datasets in the general visual domain [10, 23, 25, 28, 36, 52, 69, 74, 76, 80] reflects growing interest in understanding human visual behavior. These datasets span diverse scenarios, ranging from multi-target search tasks [28] to focused single-category search [25, 80]. Notable examples include COCO-Search18 [76] with its extensive object categories and datasets incorporating Visual Question Answering paradigms [10]. While the general domain has seen substantial progress, medical eye gaze datasets remain limited in scope. Current medical datasets concentrate primarily on 2D modalities, particularly chest X-rays, as exemplified by EGD [38] and REFLACX [6]. The absence of 3D medical eye gaze datasets represents a significant gap in the field. To address this limitation, we present the first comprehensive eye gaze dataset for CT scan analysis. CT-ScanGaze provides essential data for advancing research in volumetric medical image analysis and understanding expert visual search patterns in 3D medical contexts.

Scanpath Prediction. Early approaches to scanpath prediction primarily focused on sampling fixations from saliency maps [34, 49, 71, 73]. The field has since witnessed remarkable progress [1, 5, 9, 11–13, 15, 17, 26, 31, 35, 41, 50, 56, 58, 59, 64, 65, 72, 76–79, 81], particularly through the integration of deep neural networks [12, 17, 37, 42, 50, 59, 65, 76–78], reinforcement learning [12, 76, 77], and transformer architectures [13, 50, 59, 79]. These advances have substantially enhanced our understanding of temporal attention dynamics. However, none of previous approaches have been designed for Computed Tomography. To address this limitation, we propose a transformer-based method, CT-Searcher, specifically designed for scanpath prediction on CTs.

Table 4. Performance of CT-Searcher on different radiologists on spatial-based metrics.

Method	Saliency		
	CC ↑	KLDiv ↓	NSS ↑
Radiologist #1	0.1901±0.017	3.388±0.293	1.209±0.101
Radiologist #2	0.1503±0.014	3.912±0.331	1.072±0.089
Both	0.1706±0.016	3.645±0.312	1.142±0.095

Table 5. Ablation study on the impact of each component. **mSM** and **mMM** denote for ScanMatch and Multimatch, respectively.

Pre	3D-PE	mSM ↑	mMM ↑	SED ↓	KLDiv↓
✗	✗	0.0207±0.011	0.6034±0.031	252±23	20.958±0.902
✗	✓	0.1275±0.005	0.6883±0.022	184±18	5.194±0.523
✓	✗	0.0632±0.019	0.6562±0.019	189±12	7.526±0.616
✓	✓	**0.1318±0.009**	**0.7002±0.027**	**174±18**	**3.645±0.312**

7. Conclusion

This paper introduces CT-ScanGaze, the first public CT eye gaze dataset, along with CT-Searcher as the first CT scanpath prediction baseline. Experiments show that CT-Searcher works well and marks an important breakthrough in how we can model the way experts visually search through complex 3D medical images.

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. 8

[2] Robert Alexander, Stephen Waite, Michael A. Bruno, Elizabeth A. Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. Mandating Limits on Workload, Duty, and Speed in Radiology. *Radiology*, 304(2), 2022. 2

[3] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 6

[4] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and NoelE O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. *ECCV Workshop on Egocentric Perception, Interaction and Computing (EPIC)*, 2018. 6, 7, 18

[5] Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, and Sabine Susstrunk. TempSAL - uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[6] 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, 4, 5, 8

[7] 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. 6

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[9] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia (TMM)*, 2022. 8

[10] 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. 8

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

[12] 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. 8, 20

[13] 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, 6, 7, 8, 20

[14] Google Cloud. Speech-to-text v2 api. https://cloud.google.com/speech-to-text, 2024. Accessed: 2024-03-23. 3

[15] 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. 8

[16] 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. 6

[17] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Scanpathnet: A recurrent mixture density network for scanpath prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 8

[18] 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. 3, 6, 20

[19] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4

[20] Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021. 3, 14

[21] Trafton Drew, Karla Evans, Melissa L. H. Võ, Francine L. Jacobson, and Jeremy M. Wolfe. Informatics in Radiology: What Can You See in a Single Glance and How Might This Guide Visual Search in Medical Images? *RadioGraphics*, 33 (1), 2013. 2

[22] Trafton Drew, Melissa Le-Hoa Vo, Alex Olwal, Francine Jacobson, Steven E. Seltzer, and Jeremy M. Wolfe. Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, 13(10), 2013. 1, 2, 4

[23] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *ACM Multimedia Systems Conference (MMSys)*, 2019. 8

[24] Miguel P Eckstein, Miguel A Lago, and Craig K Abbey. The role of extra-foveal processing in 3d imaging. In *Proceedings of Spie–the International Society for Optical Engineering*. NIH Public Access, 2017. 2

[25] 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. 8

[26] Camilo Fosco, Vincent Casser, Amish Kumar Bedi, Peter O'Donovan, Aaron Hertzmann, and Zoya Bylinskii. Predicting visual importance across graphic design types. In

ACM Symposium on User Interface Software and Technology, 2020. 8

[27] 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. 6

[28] 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. 8

[29] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021. 6, 16

[30] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270, 2015. 6

[31] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 8

[32] Bulat Ibragimov and Claudia Mello-Thoms. The use of machine learning in eye tracking studies in medical imaging: A review. *IEEE journal of biomedical and health informatics*, 2024. 4

[33] 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*, pages 590–597, 2019. 16

[34] 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. 8

[35] Sen Jia and Neil D. B. Bruce. EML-NET:an expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 2020. 8

[36] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 8

[37] Yue Jiang, Luis A. Leiva, Hamed R. Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. UEyes: Understanding visual saliency across user interface types. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2023. 8

[38] 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, 4, 5, 8

[39] Naji Khosravan, Haydar Celik, Baris Turkbey, Elizabeth C Jones, Bradford Wood, and Ulas Bagci. A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning. *Medical image analysis*, 51:101–115, 2019. 4

[40] Daeun Kyung, Kyungmin Jo, Jaegul Choo, Joonseok Lee, and Edward Choi. Perspective projection-based 3d ct reconstruction from biplanar x-rays. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 5

[41] 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. 8

[42] Matthias Kümmerer, Matthias Bethge, and Thomas S. A. Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision (JoV)*, 2022. 8

[43] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods*, 45(1), 2013. 6

[44] 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. 18

[45] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6, 19

[46] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[47] Chong Ma, Lin Zhao, Yuzhong Chen, Sheng Wang, Lei Guo, Tuo Zhang, Dinggang Shen, Xi Jiang, and Tianming Liu. Eye-Gaze-Guided Vision Transformer for Rectifying Shortcut Learning. *IEEE Transactions on Medical Imaging*, 42 (11), 2023. 1

[48] Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, Dinggang Shen, Tianming Liu, and Xiang Li. Eye-gaze Guided Multi-modal Alignment for Medical Representation Learning, 2024. 1

[49] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research (VR)*, 2015. 8

[50] 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. 2, 6, 7, 8

[51] 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. 1, 4

[52] 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. 8

[53] João Pedrosa, Guilherme Aresta, João Rebelo, Eduardo Negrão, Isabel Ramos, António Cunha, and Aurélio Campilho. Lndetector: A flexible gaze characterisation collaborative platform for pulmonary nodule screening. In *XV Mediterranean Conference on Medical and Biological Engineering and Computing–MEDICON 2019: Proceedings of MEDICON 2019, September 26-28, 2019, Coimbra, Portugal*, pages 333–343. Springer, 2020. 4

[54] Peixi Peng, Wanshu Fan, Yue Shen, Wenfei Liu, Xin Yang, Qiang Zhang, Xiaopeng Wei, and Dongsheng Zhou. Eye gaze guided cross-modal alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 1

[55] 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

[56] 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. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 941–958, 2024. 8

[57] Trong-Thang Pham, Jacob Brecheisen, Carol C Wu, Hien Nguyen, Zhigang Deng, Donald Adjeroh, Gianfranco Doretto, Arabinda Choudhary, and Ngan Le. Itpctrl-ai: End-to-end interpretable and controllable artificial intelligence by modeling radiologists' intentions. *Artificial Intelligence in Medicine*, 160:103054, 2025. 4

[58] Trong Thang Pham, Tien-Phat Nguyen, Yuki Ikebe, Akash Awasthi, Zhigang Deng, Carol C. Wu, Hien Nguyen, and Ngan Le. Gazesearch: Radiology findings search benchmark. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 96–106, 2025. 8

[59] 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. 8

[60] Graciela Ramirez-Alonso, Olanda Prieto-Ordaz, Roberto López-Santillan, and Manuel Montes-Y-Gómez. Medical report generation through radiology images: an overview. *IEEE Latin America Transactions*, 20(6):986–999, 2022. 1

[61] Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. Human attention in fine-grained classification. *arXiv preprint arXiv:2111.01628*, 2021. 1

[62] 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. 16

[63] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, pages 629–633. IEEE, 2007. 3

[64] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A deep markov model of scanpath prediction for 360° images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[65] 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. 8

[66] Ashish Verma, Aupendu Kar, Krishnendu Ghosh, Sobhan Kanti Dhara, Debashis Sen, and Prabir Kumar Biswas. Artificially generated visual scanpath improves multi-label thoracic disease classification in chest x-ray images. *IEEE Transactions on Instrumentation and Measurement*, 2024. 1

[67] Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. Interpretive Error in Radiology. *American Journal of Roentgenology*, 208(4), 2017. 2

[68] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2194–2203, 2024. 1

[69] Shuo Wang, Ming Jiang, Xavier Morin, Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 2015. 8

[70] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 41(7):1688–1698, 2022. 4

[71] 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. 8

[72] Zijun Wei, Hossein Adeli, Minh Hoai, Gregory Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predict visual attention. In *NeurIPS*, 2016. 8

[73] 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. 8

[74] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014. 8

[75] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021. 8

[76] 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. 8

[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. 8

[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. 8

[79] 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)*, 2024. 2, 6, 7, 8, 18

[80] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

[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. 8

CT-ScanGaze: A Dataset and Baselines for 3D Volumetric Scanpath Modeling

Supplementary Material

Ethical Statement

This research follows all relevant ethical guidelines for medical research. All patient data used in this study was properly anonymized and de-identified following HIPAA guidelines. The radiologists who participated in the eye-tracking study provided informed consent. No personal or identifying information is included in the dataset or results. Our study aims to augment, not replace, clinical expertise and maintains the central role of human medical professionals in diagnostic decisions.

Summary

The appendix is organized as follows:

- Appendix A describes additional dataset statistics including gaze data split, number of slices distribution, and duration of data collection recording videos.
- Appendix B presents additional visualizations of our synthetic training data.
- Appendix C provides additional implementation details.
- Appendix D describes the 3D scanpath similarity metrics.
- Appendix E describes implementation details of baseline methods adapted for 3D scanpath prediction.
- Appendix F provides additional qualitative results and visualizations.
- Appendix G compares different CT Visual Encoder backbones.
- Appendix H discusses the MultiMatch simplification algorithm and analysis.
- Appendix I discuss the broader impact of our works.

A. Additional Dataset Details

A.1. Gaze Data Splits

Our dataset consists of 909 CT-gaze pairs, split into training, validation, and test sets with a ratio of 70:10:20 respectively. This translates to:

- Training set: 636 pairs
- Validation set: 90 pairs
- Test set: 183 pairs

A.2. Radiological Finding Distribution

CT-ScanGaze has a total of 9,332 findings with 60 unique finding names. The distribution of our dataset is shown in Fig. I.

A.3. Number of CT Slices

Fig. II shows the distribution of slice counts for all CT volumes. While all CT slices have a fixed resolution of 512×512 pixels in the axial plane, the number of slices varies across volumes with a total number of slices is 131,618 and 186 slices per CT in average.

A.4. Recording Duration

In our dataset, we prioritize maintaining natural workflow by allowing radiologists to read CT scans following their standard clinical practice. Table I summarizes the duration statistics of our video recordings.

Table I. Recording Duration Statistics.

	Duration (minutes)
Total recording time	4722
Average session time	5.36
Minimum session time	1.27
Maximum session time	9.68

A.5. Additional Visualization of CT-ScanGaze

We visualize a CT volume with its eye tracking data from an alternative point of view, showing all scanpaths across slices in Fig. III and temporal navigation patterns in a line chart in Fig. IV. To create Figs. III and IV, we select a CT volume such that its fixation sequence has only 48 unique slices (48 unique z values) to maintain the simplicity of the visualization. An animated video is provided, named `vis_gt.mp4`. The observed scanpath pattern demonstrates a natural progression from peripheral regions inward to areas of diagnostic significance. The timestamped report is:

```
[00:00.000 --> 00:29.960] there's a left
    upper chest pacemaker or ICD
[00:29.960 --> 00:40.240] with leads in the
     right atrium right ventricle and a
    pericardial lead along
[00:40.240 --> 00:43.240] the left
    ventricle
[00:43.240 --> 00:57.960] there are no
    enlarged axillary or supraclavicular
    lymph nodes
[00:57.960 --> 01:19.320] mildly enlarged
    paratracheal lymph nodes are present
    there's a mildly enlarged
[01:19.320 --> 01:28.320] large lymph node
    in the anterior mediastinum the left
    ventricle appears
[01:28.320 --> 01:34.680] mildly dilated
    with fatty metaplasia in the left
    ventricular apex and
```

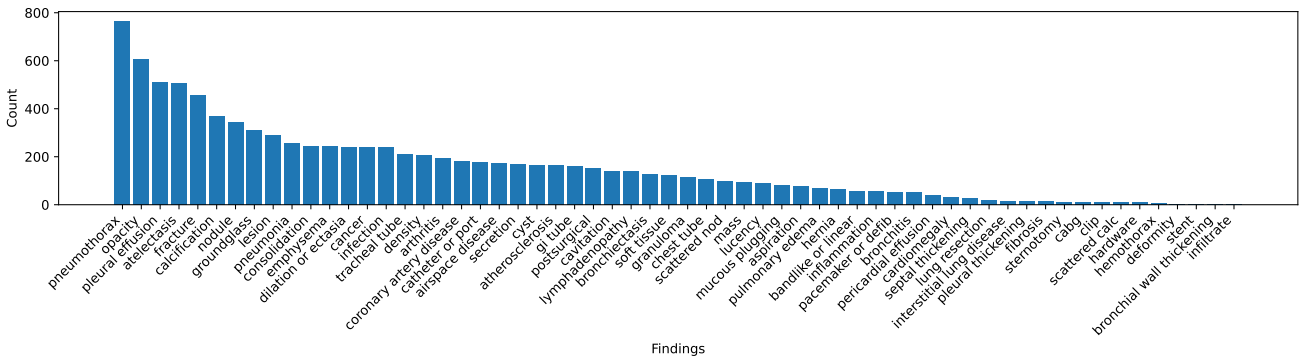

Figure I. Extracted radiological finding histogram. The y-axis represents the findings. The x-axis represents the occurrence frequency (number of samples). From SARLE [20], we extract a total of 9,332 findings with 60 unique finding names.

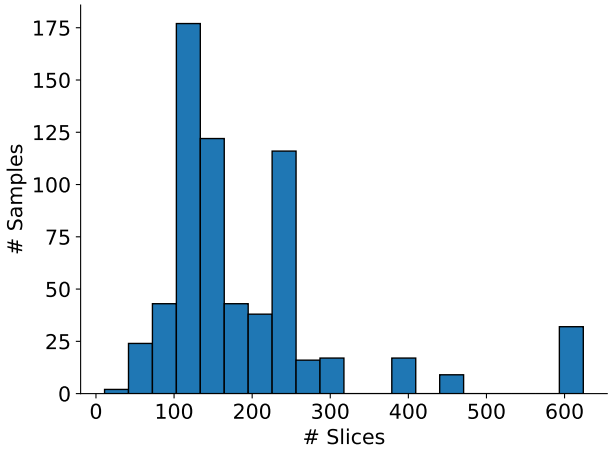

Figure II. Number of slices histogram. The y-axis represents the number of slices. The x-axis represents the occurrence frequency (number of samples). CT-ScanGaze has 131,618 slices in total and 186 slices per CT volume in average.

```
[01:34.680 --> 01:43.320] interventricular
    septum there's no pericardial effusion
    the great vessels
[01:43.320 --> 01:49.920] are normal in
    diameter there's mild aortic
    atherosclerotic calcification
[01:49.920 --> 01:55.920] there is a stent
    in the LAD
[02:01.720 --> 02:04.720] there's no
    pleural effusion
[02:04.720 --> 02:19.840] there is
    cholelithiasis a low density nodule is
    present in the left adrenal
[02:19.840 --> 02:25.800] gland likely
    representing an adenoma there's a
    calcified granuloma in the
[02:25.800 --> 02:28.800] spleen
[02:34.720 --> 02:37.720] there's no
    pericardial effusion
[02:37.720 --> 02:40.720] there's no
    pericardial effusion
[02:40.720 --> 03:06.200] a small right
    pneumothorax is present
[03:10.720 --> 03:17.720] the trachea and
    central airways are clear
[03:17.720 --> 03:31.720] a calcified
    granuloma is present in the right upper
    lobe
[03:31.720 --> 03:38.720] a small area of
    ground glass opacity is present in the
    right lower lobe
[03:38.720 --> 03:50.720] there's a right
    lower lobe nodule measuring
    approximately 15 millimeters
[04:08.720 --> 04:24.440] impression number
     one there's a small right pneumothorax
    number two a small
[04:24.440 --> 04:27.840] area of ground
    glass opacity in the peripheral right
    lower lobe is likely
[04:27.840 --> 04:37.440] infectious
    inflammatory there may be a cavitary
    component which could be the
[04:37.440 --> 04:43.240] cause of the
    right pneumothorax number three there's
    a solid right lower lobe
[04:43.240 --> 04:47.960] nodule measuring
    15 millimeters the differential includes
     infection slash
[04:47.960 --> 04:50.960] inflammation and
    malignancy
[04:54.360 --> 04:57.360] number four
[04:57.360 --> 05:00.360] mildly enlarged
    lymph nodes in the mediastinum are
    nonspecific
```

By removing the timestamps (e.g., `[00:00.000 --> 00:29.960]`), we obtain a

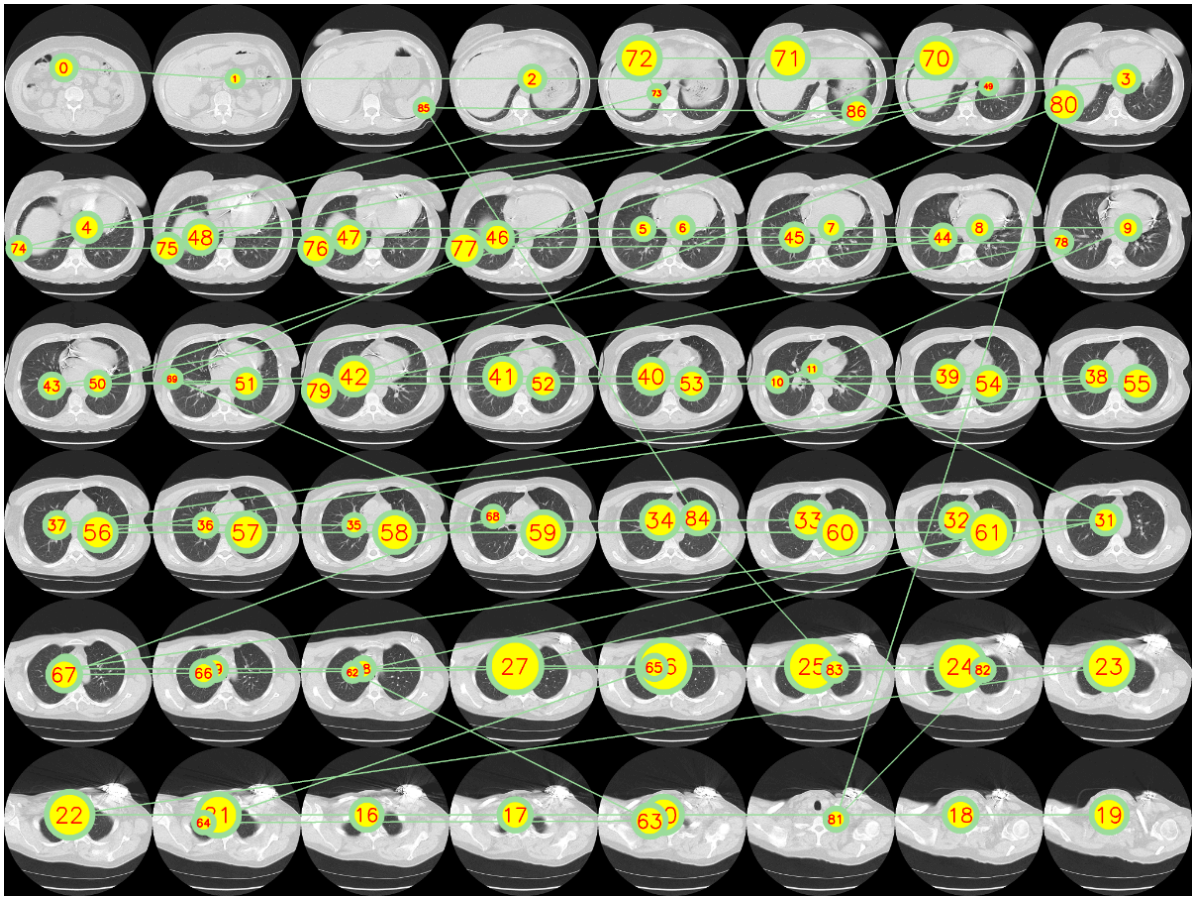

Figure III. Illustration of fixation sequence across sequential CT slices, following a left-to-right and top-to-bottom order. The fixation start from 0 in the top left corner. The numbered annotations indicate the order of fixation points. In this figure, we observe the radiologist's systematic viewing pattern: first scanning through all slices before returning to central regions for detailed examination. This navigation pattern is also demonstrated in Fig. IV. We suggest the readers to watch `vis_gt.mp4` to see the animated version of this figure.

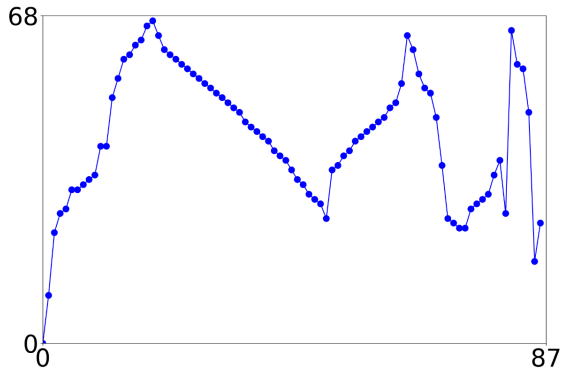

Figure IV. Visualization of temporal navigation patterns from the gaze data in Fig. III. The scanpath reveals that the radiologist follows a systematic approach: first traversing from start to end slices, then returning to central regions for detailed examination.

free-text radiology report: *There's a left upper chest pacemaker or ICD with leads in the right atrium right ventricle and a pericardial lead along the left ventricle. There are no enlarged axillary or supraclavicular lymph nodes. Mildly enlarged paratracheal lymph nodes are present. There's a mildly enlarged large lymph node in the anterior mediastinum. The left ventricle appears mildly dilated, with fatty metaplasia in the left ventricular apex and interventricular septum. There's no pericardial effusion the great vessels are normal in diameter there's mild aortic atherosclerotic calcification. There is a stent in the LAD there's no pleural effusion, there is cholelithiasis. A low density nodule is present in the left adrenal gland, likely representing an adenoma. There's a calcified granuloma in the spleen. There's no pericardial effusion. There's no pericardial effusion. A small right pneumothorax is present. The trachea and central airways are clear. A calcified granuloma is present in the right upper lobe. A small area of ground glass opacity is present in the right*

lower lobe. There's a right lower lobe nodule measuring approximately 15 millimeters. IMPRESSIONS. Number one. There's a small right pneumothorax. Number two. A small area of ground glass opacity in the peripheral right lower lobe is likely infectious inflammatory. There may be a cavitary component, which could be the cause of the right pneumothorax. Number three. There's a solid right lower lobe nodule measuring 15 millimeters, the differential includes infection slash inflammation and malignancy. Number four. Mildly enlarged lymph nodes in the mediastinum are nonspecific.

Using CheXbert [62] to extract the 13 CheXpert findings [33], we identify the following positive findings: 'Enlarged Cardiomediastinum', 'Lung Lesion', 'Pleural Effusion', and 'Support Devices'.

A.6. Example of CT-ScanGaze

We also provide one example of our data in `example_data.zip`.

- `ct_id9.nii.gz` is the CT scan.
- `finding_id9.csv` is the finding annotations.
- `fixation_id9.json` is the original fixations (without being simplified).
- `recorded_video_id9.mp4` is the recorded video session.
- `report_id9.txt` is the report created by speech to text software.

B. Additional Visualization of Synthetic Data

Fig. V illustrates samples from our synthetic dataset, comprising temporal slice navigation patterns and fixation heatmaps. Overall, the synthetic data exhibits similarities with real eye movement for 3D, particularly in temporal characteristics when transitioning slices along the depth dimension.

C. Additional Implementation Details

Due to GPU memory constraint, directly using a full CT volume as input to extract complete feature maps and train end-to-end is often not feasible. In our implementation, we follow the common practice of sliding windows and merging windows. CT-Searcher uses Swin UNETR [29] encoder with $96\times96\times96$ input window. The output shape is reduced to $3\times3\times3$ with feature dimension $C=768$. With a downsampling ratio $r=32$, we merge all features into a single feature map with shape $(W',H',D')=(W/r,H/r,D/r)$. For technical convenience, we interpolate all feature maps to a standardized size of $16\times16\times16$ with $C=768$ channels, ensuring uniformity across varying feature map shapes. This standardization is reasonable since our CT scans have fixed dimensions of 512×512 in height and width (with only depth D varying), and the feature size of $16\times16\times16\times768$ keeps the model within our GPU memory (48 GB VRAM).

D. 3D Scanpath Metrics

Due to the complexity of 3D scanpath metrics, we describe them at a high level and point out the modifications from 2D to 3D. For detailed implementation, we encourage readers to examine the source code directly:

- `visual_attention_metrics.py`: Contains implementations of:
  - Saliency metrics Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), Kullback-Leibler divergence (KLDiv).
  - String-edit-distance (SED) metric.
- `scanmatch3d.py`: Contains the 3D-adapted version of ScanMatch.
- `multimatch_3dgaze.py`: Contains the 3D-adapted version of MultiMatch.

D.1. Saliency Metrics

All three saliency metrics, CC (Correlation Coefficient), NSS (Normalized Scanpath Saliency), and KLDiv (Kullback-Leibler Divergence), are based on heatmaps and can be used directly without modification. The CC metric is defined as:

$$\hat{S}=\frac{S-\mu_S}{\sigma_S}$$
$$\hat{G}=\frac{G-\mu_G}{\sigma_G}$$
$$CC(S,G)=\frac{\sum_{i,j,k}\hat{S}_{ijk}\hat{G}_{ijk}}{\sqrt{\sum_{i,j,k}\hat{S}^2_{ijk}\sum_{i,j,k}\hat{G}^2_{ijk}}} \tag{11}$$

where $S\in[0,1]^{H\times W\times D}$ is the saliency map, $G\in\{0,1\}^{H\times W\times D}$ is the ground truth fixation map, and μ and σ are mean and standard deviation. Given the fixation sequence $\{(x_l,y_l,z_l)\}_{l=1}^N$, where N is the fixation length, the ground truth map is defined as:

$$G_{ijk}=\begin{cases}1 & \text{if } (i,j,k)\in\{(x_l,y_l,z_l)\}_{l=1}^N\\ 0 & \text{otherwise}\end{cases} \tag{12}$$

Higher CC scores indicate better matching between sequences, with an upper bound of 1.0.

The NSS metric is defined as:

$$\tilde{S}=\begin{cases}\frac{S}{\max(S)} & \text{if } \max(S)\neq0\\ S & \text{otherwise}\end{cases}$$
$$\hat{S}=\begin{cases}\frac{\tilde{S}-\mu_{\tilde{S}}}{\sigma_{\tilde{S}}} & \text{if } \sigma_{\tilde{S}}\neq0\\ \tilde{S} & \text{otherwise}\end{cases} \tag{13}$$
$$NSS(\hat{S},F)=\frac{1}{N}\sum_{i,j,k}\hat{S}_{ijk}G_{ijk}$$

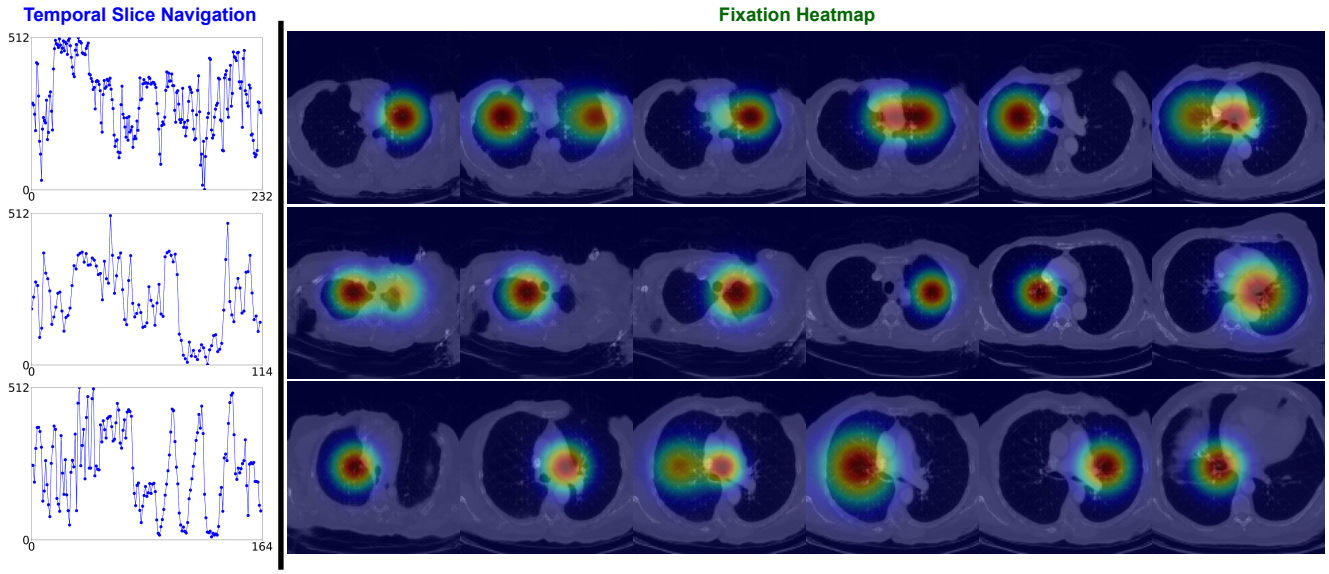

Figure V. Visualization of our synthetic training data. The Temporal Slice Navigation demonstrates the temporal transition of slice. The Fixation Heatmap column displays representative CT slices with corresponding fixation heatmaps.

where $\hat{S}$ is the normalized saliency map. Higher NSS scores indicate better matching between sequences, with an upper bound that depends on the ground truth.

The KLDiv metric is defined as:

$$\begin{aligned} \tilde{S}_{ijk} &= \frac{S_{ijk}}{\sum_{i,j,k} S_{ijk}} \\ \tilde{G}_{ijk} &= \frac{G_{ijk}}{\sum_{i,j,k} G_{ijk}} \\ KLDiv(S,G) &= \sum_{i,j,k} \tilde{G}_{ijk} \log\left(\epsilon + \frac{\tilde{G}_{ijk}}{\tilde{S}_{ijk} + \epsilon}\right) \end{aligned} \tag{14}$$

where $\epsilon = 2.2204 \times 10^{-16}$ is a small constant to prevent divide-by-zero and log-zero. Lower KLDiv scores indicate better matching between sequences, with a lower bound of 0.0.

D.2. String-edit-distance

String-edit-distance (SED) has two main steps:

1. Converts gaze sequences into strings:
   a) Dividing the 3D volume into discrete cells. In the original 2D version, this step divides the image into patches.
   b) Assigning unique characters to each cell.
   c) Mapping fixation points to these characters in sequence.
2. Compares two sequences using Levenshtein distance by counting minimum number of operations (insertions, deletions, substitutions).

In our 3D adapted SED, we change the step 1.a from 2D into 3D, the other steps are left as is. Lower SED scores indicate better matching between sequences, with a lower bound of 0.0.

D.3. ScanMatch

Calculating ScanMatch (SM) score between predicted and ground truth fixations consists of 3 main steps:

1. Convert fixation sequences into letter strings. This step is similar to the first step of SED. In addition, when considering duration (ScanMatch w/ Dur.), each character is repeated n times, where n is the duration in milliseconds. This repetition is not performed when duration is not considered (ScanMatch w/o Dur.).
2. Create a substitution matrix with scores for all possible letter pairs. The original score function uses 2D Euclidean distance, which we extend to 3D Euclidean distance.
3. Sequence comparison:
   a) Create comparison matrix:
      - Columns: letters from first sequence
      - Rows: letters from second sequence
      - Cell values: costs from substitution matrix
   b) Apply Needleman-Wunsch algorithm to find optimal alignment path
   c) Calculate normalized similarity score (0-1 scale)

We adapt to 3D by modifying both step 1 and the substitution matrix calculation at step 2, replacing 2D Euclidean distance with 3D Euclidean distance. Higher SM scores indicate better matching between sequences, with an upper bound of 1.0.

D.4. MultiMatch

Different from ScanMatch and SED, MultiMatch (MM) measures scanpath similarity regarding shape, direction, length, position, and duration. Higher MM scores indicate better sequence matching, with an upper bound of 1.0 for all aspects. Given the predicted fixations $\{(\hat{x}_l, \hat{y}_l, \hat{z}_l, \hat{t}_l)\}_{l=1}^{N}$ and ground truth fixations $\{(x_l, y_l, z_l, t_l)\}_{l=1}^{N}$, we calculate MultiMatch scores with 3 main steps:

1. Temporal alignment:
   a) Calculate how similar each element i in one scanpath is compared to each element j in the other scanpath based on a similarity metric. Collect all pairs (i, j) to create a similarity matrix $M(i, j)$ between elements.
   b) From $M(i, j)$, build adjacency matrix A with connection weights like a graph.
   c) Find the shortest path from i to j using Dijkstra's algorithm.
   d) Align scanpaths along shortest path.
2. For every align pair of fixation (i, j), we compute similarity across five dimensions:
   a) **Vector (shape):** shape difference between fixation vectors $(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i) - (x_j, y_j, z_j, t_j)$.
   b) **Direction:** difference in direction (angle) between fixation vectors. We measure angles using spherical coordinates in 3D, analogous to the original authors' use of polar coordinates in 2D.
   c) **Length:** difference in amplitude (length) between fixation vectors $\left|(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i) - (x_j, y_j, z_j, t_j)\right|$,
   d) **Position:**: 3D Euclidean distance between fixations.
   e) **Duration:**: difference in duration between fixations.
3. Score normalization:
   a) Vector, Length, and Position scores are normalized by volume diagonal.
   b) Direction is normalized by π.
   c) Duration is normalized by maximum duration.

E. 3D Scanpath Prediction Baselines

E.1. PathGAN

Similar to original PathGAN [4], our CT-adapted PathGAN architecture has two major components: a Discriminator D and a Generator G. The Generator takes a CT volume V as input and produces a fixation sequence $G(V) = \{(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i)\}_{i=1}^{N}$. The Discriminator aims to assign low scores to N predicted fixations $(\hat{x}_i, \hat{y}_i, \hat{z}_i, \hat{t}_i)\}_{i=1}^{N}$ and high scores to N_{gt} ground truth fixations $gt = \{(x_i, y_i, z_i, t_i)\}_{i=1}^{N_{gt}}$, taking both the fixation sequence and CT volume features as input. The PathGAN architecture is illustrated in Fig. VI. Note that we share a frozen Swin UNETR module between G and D during training while optimizing other modules.

For the loss functions, we maintain PathGAN's default implementation using two main components: conditional

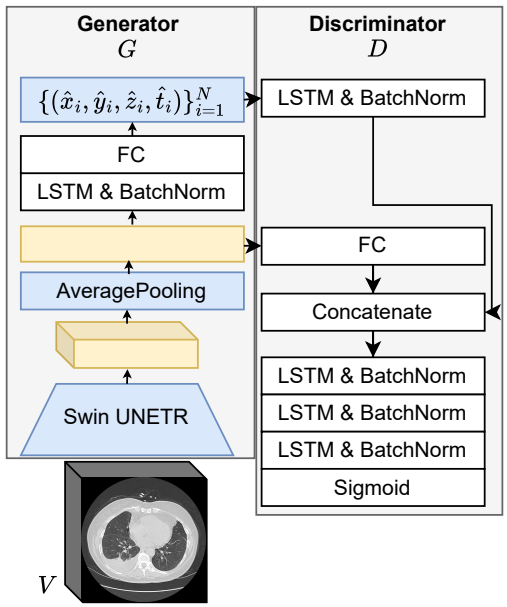

Figure VI. Our adapted version of PathGAN for CT scanpath prediction maintains the core architecture while altering components: the visual encoder module to Swin UNETR, Average Pooling to 3D, and predicted fixations (highlighted in blue).

GAN loss and L^2 loss between ground truth and predicted fixations. Specifically, the conditional GAN loss is defined as:

$$\mathcal{L}_{\text{cGAN}} = \mathbb{E}_{V,gt}[\log D(V, gt)] + \mathbb{E}_{V}[\log(1 - D(V, G(V))], \tag{15}$$

The L2 loss is defined as:

$$\mathcal{L}_{L^2} = \mathbb{E}_{V,gt}\left[\|gt - G(V)\|^2\right] \tag{16}$$

The final formulation of the loss function for the generator during adversarial training is:

$$\mathcal{L} = \mathcal{L}_{\text{cGAN}} + \alpha\mathcal{L}_{L^2} \tag{17}$$

Following the original PathGAN implementation, we set the hyperparameter $\alpha = 0.05$.

E.2. HAT

The detailed adapted HAT [79] architecture is shown in Fig. VII. Similar to HAT's FPN visual encoder [44] that generates features at both bottleneck and high-resolution levels, we extract features from two Swin UNETR layers: $P1 \in \mathbb{R}^{H/32 \times W/32 \times D/32 \times C}$ from the bottleneck layer as low resolution feature and $P4 \in \mathbb{R}^{H/4 \times W/4 \times D/4 \times C_4}$ from the 4^{th} layer of Swin UNETR's decoder as high resolution feature. We extend HAT's original loss to handle 3D fixation maps and use a single query as the class query. Note that Fig. VII shows one-step prediction because HAT predicts fixations step by step, with the working memory being updated after each fixation heatmap prediction [79].

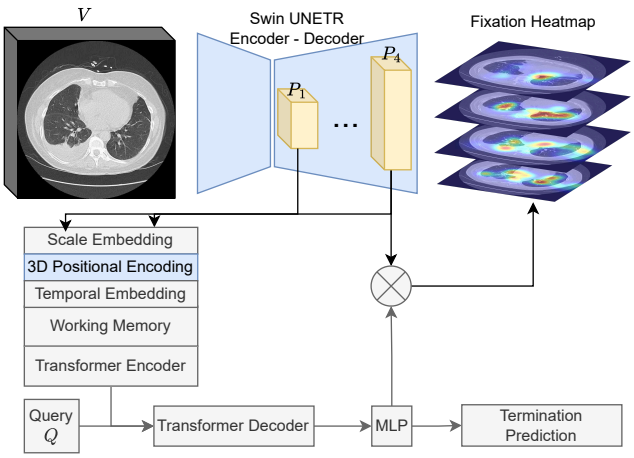

Figure VII. Our adapted version of HAT for CT scanpath prediction maintains most of the original architecture while modifying key components: the Visual Encoder (Swin UNETR), 3D Position Encoding, and prediction output (highlighted in blue). $\otimes$ is matrix multiplication.

Then, given predicted fixation heatmaps $\hat{Y} \in [0,1]^{H\times W\times D}$ and termination probabilities $\hat{\tau} \in [0,1]$, we compute the loss at each step i:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\text{fix}}\left(\hat{Y}_i, Y_i\right) + \mathcal{L}_{\text{term}}\left(\hat{\tau}_i, \tau_i\right) \tag{18}$$

where $Y_i \in [0,1]^{H\times W\times D}$ represents the ground-truth 3D fixation heatmap, $\tau_i \in \{0,1\}$ is the termination label, and N is the length of the fixation sequence. We generate Y by applying a Gaussian kernel with sigma equal to 1 visual angle to the ground-truth fixation map. The fixation loss $\mathcal{L}_{\text{fix}}$ is a volumetric focal loss:

$$\mathcal{L}_{\text{fix}} = \frac{-1}{HWD}\sum_{i,j,k}\begin{cases} \left(1-\hat{Y}_{ijk}\right)^{\alpha}\log\left(\hat{Y}_{ijk}\right) & \text{if } Y_{ijk}=1 \\ \left(1-Y_{ijk}\right)^{\beta}\left(\hat{Y}_{ijk}\right)^{\alpha} & \\ \log\left(1-\hat{Y}_{ijk}\right) & \text{otherwise} \end{cases} \tag{19}$$

with $\alpha = 2$ and $\beta = 4$. And the termination loss $\mathcal{L}_{\text{term}}$ is a binary cross entropy loss:

$$\mathcal{L}_{\text{term}} = -\tau\log\left(\hat{\tau}_i\right) - (1-\tau)\log\left(1-\hat{\tau}_i\right) \tag{20}$$

E.3. Gazeformer

Our adapted version of Gazeformer for CT scanpath prediction is illustrated in Fig. VIII. Besides adapting the architecture to our task, we also extend the loss function to 3D. The total loss function is defined as:

$$\mathcal{L} = \left(\mathcal{L}_{xyzt} + \mathcal{L}_{val}\right) \tag{21}$$

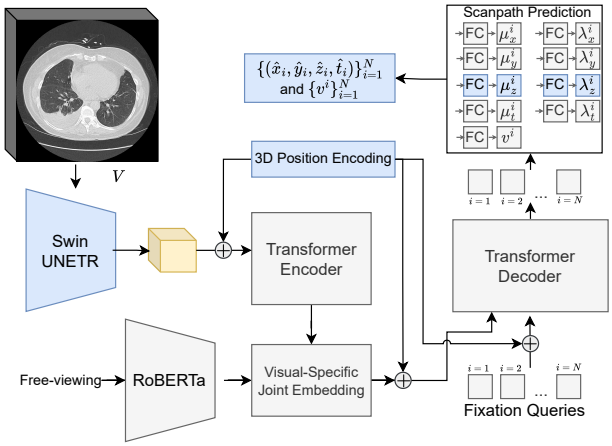

Figure VIII. Our adapted version of Gazeformer for CT scanpath prediction maintains most of the original architecture while modifying three key components (highlighted in blue): replacing the visual encoder with Swin UNETR, incorporating 3D Position Encoding, and adding an extra prediction head for z-coordinate distribution in the Scanpath Prediction module. We use 'freeview' as input for RoBERTa [45] and modify the Scanpath Prediction module to predict three separate branches for x, y, and z coordinates.

where the coordinate regression loss is:

$$\mathcal{L}_{xyzt} = \frac{1}{N_{gt}}\sum_{i=1}^{N_{gt}}\left(|x_i-\hat{x}_i| + |y_i-\hat{y}_i| + |z_i-\hat{z}_i| + \left|t_i-\hat{t}_i\right|\right) \tag{22}$$

and the validity prediction loss is:

$$\mathcal{L}_{val} = -\frac{1}{N}\sum_{i=1}^{N}\left(\hat{v}_i\log v_i + (1-\hat{v}_i)\log\left(1-v_i\right)\right) \tag{23}$$

Here, $\left\{\left(\hat{x}_i,\hat{y}_i,\hat{z}_i,\hat{t}_i\right)\right\}_{i=1}^{N}$ represents the predicted scanpath and N is the maximum predicted scanpath length. N_{gt} denotes the length of the ground truth scanpath $\{(x_i,y_i,z_i,t_i)\}_{i=1}^{N_{gt}}$. The binary scalar v_i indicates whether the i^{th} token in the ground truth fixation is a valid fixation or padding, while $\hat{v}_i$ represents our model's predicted probability of token validity.

E.4. GazeformerISP

Our adapted version of GazeformerISP for CT scanpath prediction replaces the original encoder with Swin UNETR encoder's bottleneck features, 3D Position Encoding, prediction output and extends the loss function to handle 3D fixation maps. The architecture is illustrated in Fig. IX. GazeformerISP predicts and computes loss on 3D fixation maps to represent 3D coordinates. The objective jointly optimizes the fixation map $\hat{Y}_i$ and duration $\hat{t}_i$:

$$\mathcal{L} = -\sum_{i=1}^{N} Y_t\log\hat{Y}_t + \sum_{t=1}^{N}\left|t_i-\hat{t}_i\right| \tag{24}$$

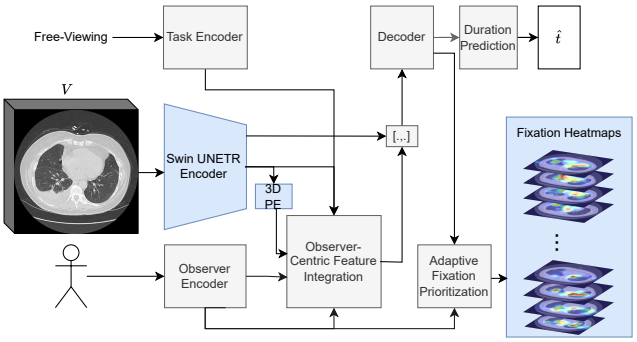

Figure IX. Our adapted version of GazeformerISP for CT scanpath prediction maintains most of the original architecture while modifying three key components (highlighted in blue): the input features, 3D Position Encoding, and prediction output. Here, [..,] denotes the concatenation operator, and '3D PE' represents 3D Positional Encoding. Similar to HAT's freeview mode, we employ a single query embedding to predict fixations autoregressively.

where N is the maximum length of fixations, Y_t and t_i represent the ground-truth fixation maps and fixation duration, respectively. Finally, we train GazeformerISP with Self-Critical Sequence Training (SCST) set up using ScanMatch as reward, and the Consistency Divergence loss as originally described in [12, 13].

F. Additional Qualitative Results

Fig. X presents an additional comparison of the temporal slice navigation and fixation heatmaps across multiple CT slices between CT-Searcher with state-of-the-art scanpath prediction methods. CT-Searcher outperforms others by capturing a balance between realism and variability, avoiding excessive noise or oversimplification in the temporal slice navigation comparison. Additionally, CT-Searcher achieves more visually faithful heatmaps compared to the ground-truth, outperforming other approaches in detail and accuracy. For baseline methods, we observe several limitations. Some methods produce no heatmaps (N/A) in certain positions, showing their inability to generate meaningful outputs. Similar to PathGAN, Gazeformer covers limited CT slices, indicating a constraint in handling 3D fixation tasks. Both GazeformerISP and HAT can produce heatmaps for most CT slices, however their temporal slice navigation appears noisy and inconsistent, deviating from the ground-truth pattern. In conclusion, our method outperforms other approaches and mimics ground truth scanpath in both temporal slice navigation and fixation heatmap generation.

G. Comparison of CT Backbone Architectures

Tab. II demonstrates the comparative performance of CT-ViT and Swin UNETR backbones as CT-Searcher's Visual Encoder across multiple scanpath similarity metrics:

Table II. Ablation: Comparison of backbone architectures across scanpath metrics. Arrows (↑/↓) indicate whether higher or lower scores are better. Bold values indicate the best performance.

Visual Encoder	SM ↑	MM ↑	SED ↓	KLDiv↓
CT-ViT	0.1287	0.6934	193	3.665
Swin UNETR	**0.1318**	**0.7002**	**174**	**3.645**

Table III. Fixation count comparison between the original gaze data and the simplified gaze data.

Version	Original	Simplified	Reduction (%)
Number of Fixations	2,234,920	954,311	57.3%

Table IV. MultiMatch similarity scores between the original gaze data and the simplified gaze data.

Dimension	Vector	Direction	Length	Position	Average
Score	0.993	0.853	0.989	0.944	0.945

ScanMatch (SM), MultiMatch (MM), String-Edit Distance (SED), and Kullback-Leibler Divergence (KLDiv). We freeze both backbones during training. Tab. II shows that Swin UNETR outperforms CT-ViT across all metrics, achieving higher scores in pattern-based measures (SM: 0.1318 vs. 0.1287, MM: 0.7002 vs. 0.6934) and lower values in distance-based metrics (SED: 174 vs. 193, KLDiv: 3.645 vs. 3.665). Based on the empirical results, we adopt Swin UNETR as our Visual Encoder.

H. MultiMatch Simplification Analysis

Due to the original gaze data containing scanpaths with numerous fixations that are dense and complex with sequences averaging 543 fixations and reaching up to 2,708 fixations per CT, we employ the simplification algorithm to make this data more manageable while preserving essential gaze patterns, from the MultiMatch toolbox [18] with default settings: an angular threshold of 45° and an amplitude threshold of 10% of the volume resolution diagonal. This reduces sequences to an average of 222 fixations with a maximum of 1,507 fixations. Both original and simplified versions will be made available.

To demonstrate the effectiveness of simplification, Figs. XI to XIII presents a comparative illustration between original and simplified scanpaths. While radiologists' eye movements on a single CT slice generally focus on the image center, movement along the slice dimension exhibits more complexity. Fig. XI reveals continuous and intricate radiologist navigation through depth. Nevertheless, the MM simplification approach introduces only minor changes to the movement landscape while preserving the overall pattern. This consistency in pattern preservation is also evident along the x-axis (Fig. XII) and y-axis (Fig. XIII). In sum-

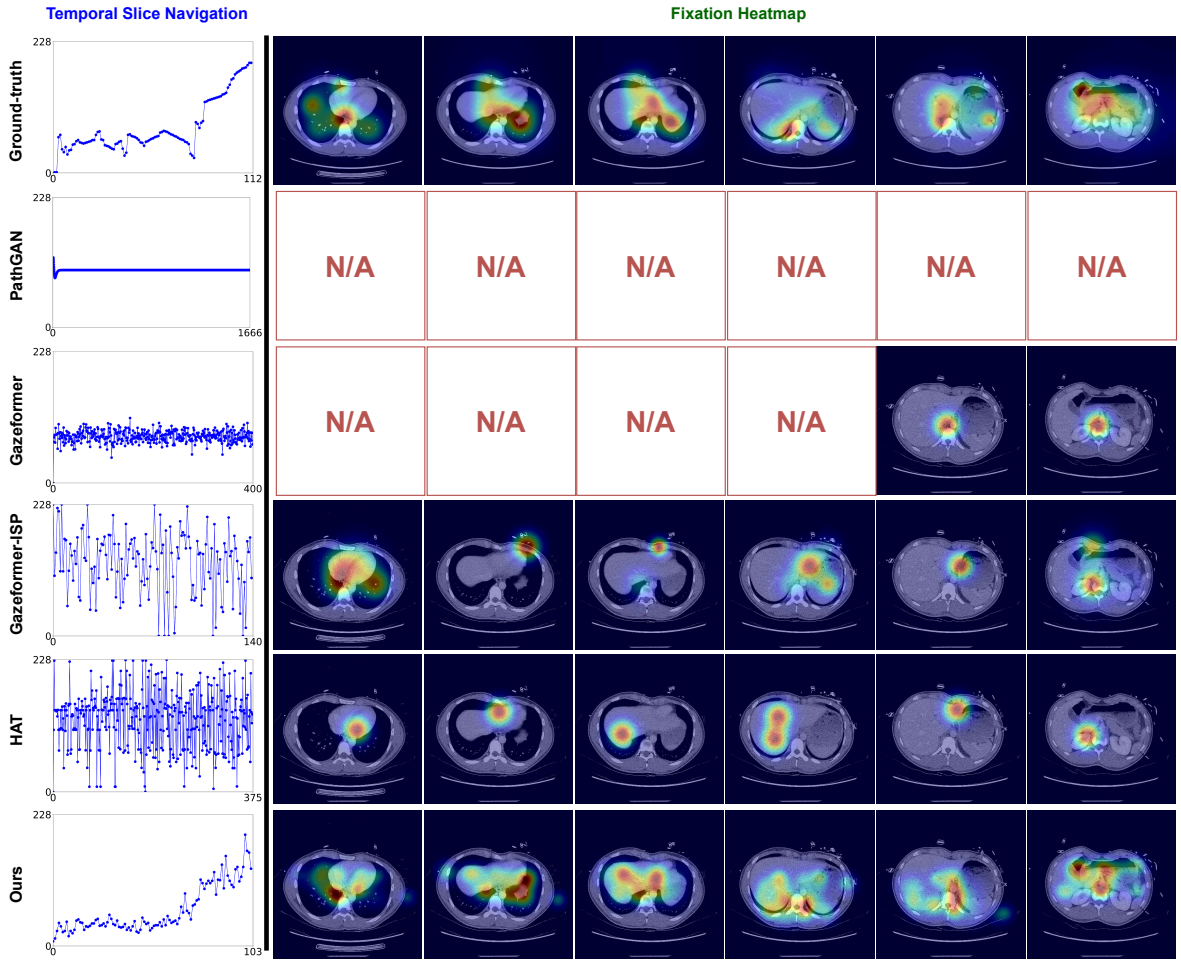

Figure X. Additional qualitative results between CT-Searcher and state-of-the-art scanpath prediction methods. N/A in a particular slice position (column) means that the corresponding model (row) fails to predict scanpath for that slice, thus no heatmap can be created. The heatmap images in the Fixation Heatmap column show the eye gaze fixation patterns across different CT slices. The left columns show Temporal Slice Navigation patterns, illustrating how scanpath traverses through different slices over time.

mary, the line charts demonstrate that while significantly reducing the number of points (Tab. III), the simplification process maintains the essential scanpath characteristics, as evidenced by minimal changes in MultiMatch similarity scores across all spatial dimensions (Tab. IV).

I. Discussion

Our contributions establish a foundation for volumetric scanpath modeling. CT-ScanGaze benefits the research communities in several ways. First, it provides a benchmark specifically designed for 3D scanpath prediction in medical imaging, addressing limitations of 2D-focused datasets. Second, by capturing radiologists' visual attention patterns during diagnosis, it enables research on the relationship between visual search behavior and diagnostic reasoning, advancing explainable AI in heathcare. Furthermore, CT-ScanGaze enables several research directions: analyzing expert vs. novice radiologist gaze patterns, developing generalizable 3D attention models, creating radiology training protocols based on expert viewing patterns, and designing human-AI collaborative systems that leverage natural viewing behaviors. CT-Searcher demonstrates the feasibility of modeling expert visual behavior in CT interpretation. While our current implementation focuses on CT scans, the approach could extend to other domains requiring 3D visualization expertise, such as geological analysis or industrial CT inspection. Future work should address current limitations, including expanding the dataset to encompass more diverse pathologies, developing more interpretable models that can explain predicted attention patterns, and evaluating the clinical impact of these systems on diagnostic accuracy and efficiency in real-world settings.

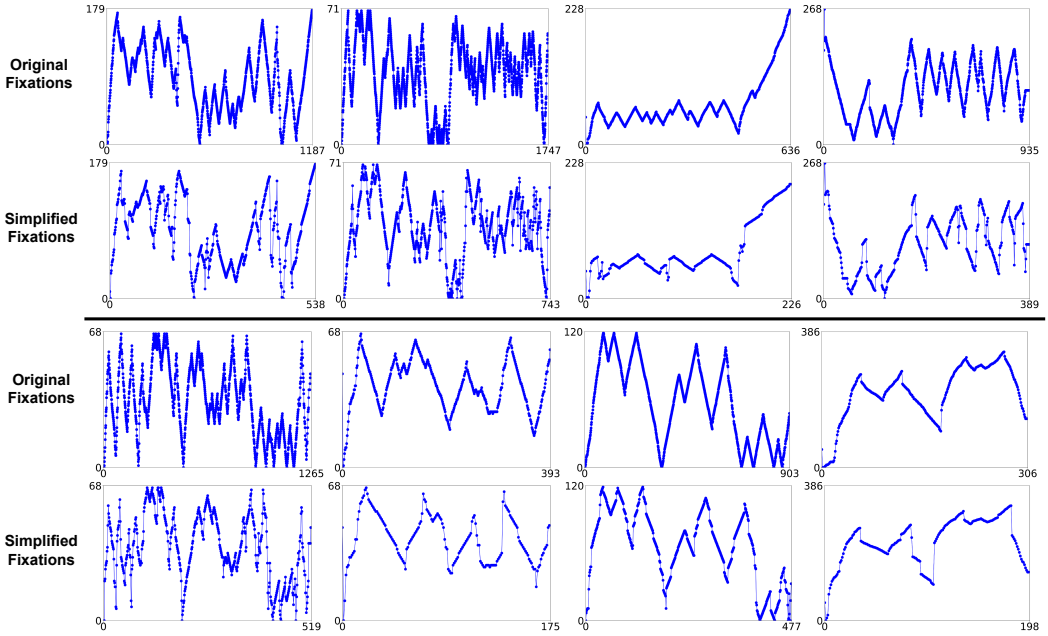

Figure XI. The effect of MM simplification algorithm on the z (slice) dimension.

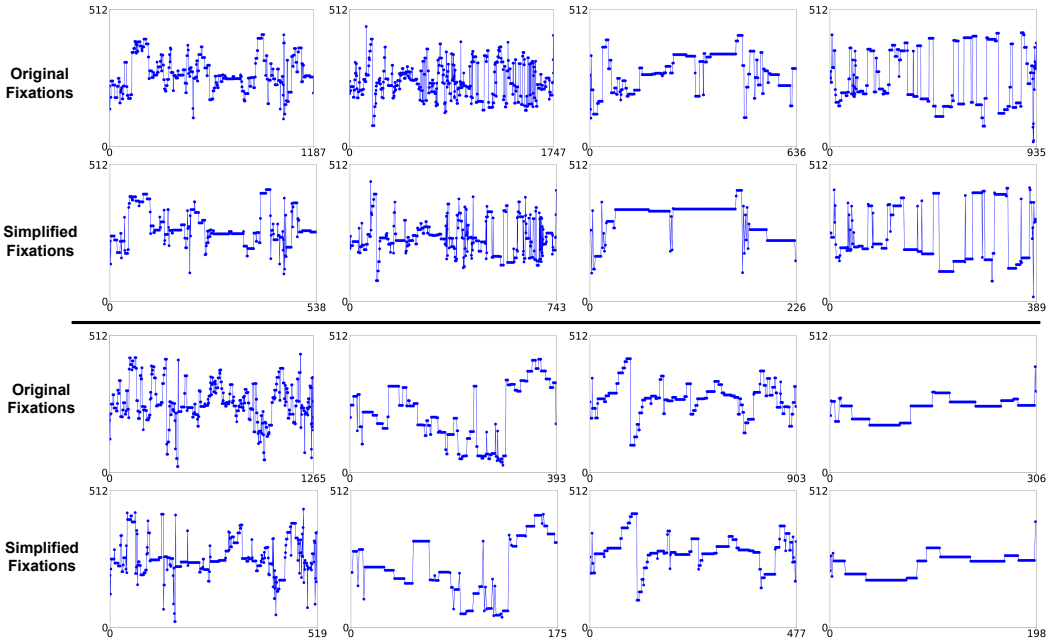

Figure XII. The effect of MM simplification algorithm on the x (width) dimension.

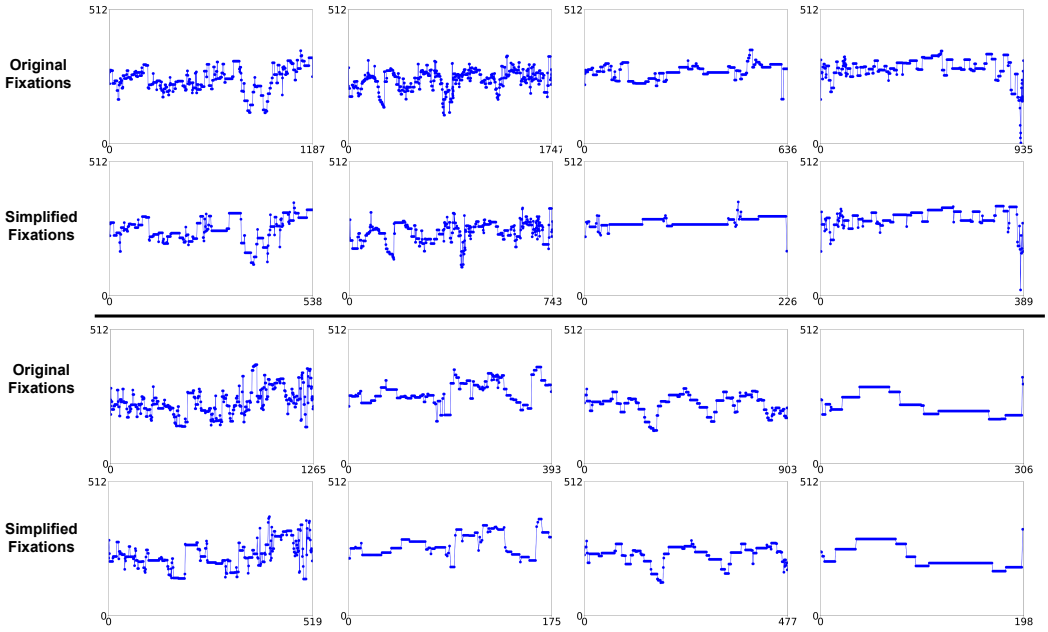

Figure XIII. The effect of MM simplification algorithm on the y (height) dimension.