HiSTAR: Hierarchical Semantic-aware Spatio-Temporal Decoding with Vision-Language Models for Radiology Scanpath Prediction

Anonymous authors
Paper under double-blind review

Abstract

Radiology scanpath prediction aims to model the dynamic visual search behavior of radiologists during diagnostic reasoning. Radiological abnormalities typically exhibit low contrast and subtle visual manifestations, requiring integration of anatomical context with localized abnormal findings during visual search. Consequently, accurate scanpath prediction requires jointly modeling abnormality localization and the temporal evolution of visual attention, posing substantial challenges for existing methods. Recent medical vision-language models provide rich clinical context that complements purely visual representations. Motivated by these advances, we propose HiSTAR, a Hierarchical Semantic-aware Spatio-Temporal Decoding framework for radiology scanpath prediction. With HiSTAR, hierarchical semantic adaptation incorporates global and granularity clinical semantics to construct image-specific diagnostic search conditions and semantic-aware regional representations. Spatial-guided temporal decoding estimates a task-relevant spatial attention distribution as a prior for progressive trajectory generation. Finally, geometry-attention consistency further regularizes attention allocation and inter-fixation geometry to promote diagnostically relevant and spatially coherent scanpaths. HiSTAR achieves state-of-the-art performance on the GazeSearch benchmark, with a ScanMatch score of 0.3942. Downstream disease classification on SIIM-ACR and TB-MOUSE further demonstrates that HiSTAR-predicted gaze provides effective task-relevant information for medical image analysis. Code is available at https://anonymous.4open.science/r/HiSTAR.

1 INTRODUCTION

Radiology scanpath prediction seeks to model and predict the visual search behavior of radiologists performing specific diagnostic tasks (Pham et al., 2025c). Given a radiological image and a diagnostic target, the objective is to predict a scanpath composed of sequential fixation locations and their durations. The trajectories characterize how radiologists continuously allocate and shift visual attention under the joint influence of diagnostic objectives, image evidence, and medical knowledge (Wu et al., 2026a; Nan et al., 2025). Modeling radiological visual search can deepen the understanding of radiologists' perceptual and cognitive processes, provide process-level evidence for computer-aided diagnosis, offer expert attention patterns for radiology training, and support the development of interactive medical systems (Zhu et al., 2026; Pham et al., 2025b; Awasthi et al., 2025).

However, radiology scanpath prediction differs fundamentally from target search in natural scenes. Searching for a cup in a natural image generally relies on well-defined object boundaries and appearance cues, whereas identifying consolidation on a chest radiograph requires reasoning over anatomical relationships and subtle radiographic findings (Pham et al., 2025a). Consequently, an effective scanpath prediction model needs not only to localize abnormal regions, but also to model the temporal relationships of visual attention across these regions (Lou et al., 2025; Wu et al., 2026b).

Previous studies have progressively incorporated domain knowledge to address these challenges. ChestSearch adopts medical-domain pretraining and multi-scale features to learn clinically relevant representations (Pham et al., 2025c), while LogitGaze-Med further integrates patch-level categorical semantics with visual features for joint modeling (Lvov & Pershin, 2025b). Nevertheless, existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

semantic information remains largely limited to image-level descriptions or patch-level categorical cues, providing insufficient characterization of fine-grained clinical attributes and lesion-specific findings. Meanwhile, scanpath prediction requires joint modeling of spatial localization and temporal dependencies; coupling the two within a single prediction stage increases generation complexity and can compromise trajectory stability, while the geometric structure of inter-fixation transitions also remains insufficiently modeled (Palmer et al., 2026).

To address the above limitations, we propose **HiSTAR**, a Hierarchical Semantic-aware Spatio-Temporal Decoding framework for radiology scanpath prediction. HiSTAR leverages medical knowledge in vision-language models (VLMs) to identify diagnostically relevant spatial regions before decoding the temporal search trajectory, supplementing visual evidence with semantic context while simplifying scanpath prediction through spatial and temporal modeling. Specifically, we introduce Hierarchical Semantic Adaptation (HSA) to incorporate clinical semantics at two complementary levels. Global context adapts the target finding into an image-specific search condition, while granularity semantics enrich visual features with anatomical and abnormality-related information. Together, the hierarchical semantics provide clinically informed cues for subsequent scanpath prediction (Chen et al., 2026). We further propose Spatial-guided Temporal Decoding (STD) to model scanpath prediction as spatial attention localization and temporal trajectory generation. A task-relevant spatial attention map is first estimated to identify where to look, and the resulting spatial prior then guides progressive trajectory decoding to determine how attention evolves. Finally, Geometry-Attention Consistency (GAC) regularizes both inter-fixation geometry and attention allocation, encouraging more coherent and diagnostically relevant scanpaths.

We conduct experiments on the standard benchmark GazeSearch to evaluate HiSTAR for radiology scanpath prediction (Pham et al., 2025c). HiSTAR consistently outperforms existing methods across multiple scanpath similarity metrics, achieving ScanMatch scores of 0.3942 and 0.3664 without and with duration, respectively. Downstream classification on SIIM-ACR (Zawacki et al., 2019) and the self-collected TB-MOUSE further demonstrates that HiSTAR-predicted gaze improves diagnostic performance, indicating that the generated scanpaths preserve disease-relevant information. Our contributions can be summarized as follows:

- We propose **HiSTAR**, a Hierarchical Semantic-aware Spatio-Temporal Decoding framework for radiology scanpath prediction. HiSTAR leverages medical knowledge from a VLM through **hierarchical semantic adaptation**, combining global clinical context with granularity semantics to provide image-specific semantic cues for scanpath prediction.
- We develop **spatial-guided temporal decoding** to simplify scanpath prediction by separating task-relevant spatial localization from temporal trajectory generation. We further introduce **geometry-attention consistency** to constrain both attention allocation and inter-fixation transitions.
- Experiments on GazeSearch demonstrate that HiSTAR outperforms existing methods across multiple scanpath similarity metrics. Downstream classification on SIIM-ACR and TB-MOUSE further shows that HiSTAR-predicted gaze preserves disease-relevant information beneficial for medical image analysis.

2 RELATED WORK

2.1 SCANPATH PREDICTION

Human visual perception is inherently dynamic, with observers acquiring local evidence through fixations and shifting attention via saccades (Murlidaran & Eckstein, 2025; Ghosh et al., 2024). Since fixation locations and transition patterns are influenced by image content, task objectives, target semantics, and fixation history, scanpath prediction aims to model sequential visual search by predicting fixation locations and durations under these conditions (Agrawal et al., 2026; Mondal et al., 2025). Early methods often formulated this task as sequential decision making; for example, ChenLSTM uses reinforcement learning to generate scanpaths for visual question answering (Chen et al., 2021a; 2024). More recent approaches increasingly adopt Transformer-based sequence modeling. Gazeformer jointly models visual features and target semantics for target-driven scanpath prediction (Mondal et al., 2023), while HAT combines top-down and bottom-up attention with dense prediction to reduce information loss from spatial discretization (Yang et al., 2024). In addition, Xue

et al. (2025) model inter-observer variability from a small number of individual fixation samples, enabling personalized scanpath generation.

In radiological imaging, GazeSearch establishes a benchmark for chest X-ray visual search and introduces ChestSearch, which incorporates medical knowledge through domain-specific pretraining and progressively decodes scanpaths (Pham et al., 2025c). LogitGaze-Med further integrates VLM-derived patch-level categorical semantics with visual features to improve scanpath prediction (Lvov & Pershin, 2025b). CT-ScanGaze extends scanpath prediction to 3D CT volumes with CT-Searcher (Pham et al., 2025a). In addition, Chen et al. (2026) introduce image-level semantics extracted by VLMs to bridge human visual attention and vision-language representations. Despite these advances, existing methods mainly rely on image-level descriptions or patch-level categorical semantics, leaving finer-grained clinical attributes and lesion-specific findings underexplored. Meanwhile, spatial and temporal modeling are typically coupled, with limited characterization of inter-fixation transition geometry.

2.2 Vision-Language Models in Medical Imaging

In recent years, large-scale VLMs have introduced high-level semantic knowledge into visual analysis. Pretrained models such as contrastive language-image pre-Training (CLIP) have demonstrated strong generalization across a wide range of natural-image tasks (Radford et al., 2021). In the medical domain, MedCLIP, BiomedCLIP, and PLIP align medical images with domain-specific language, enabling richer semantic interpretation of medical images and improving the recognition of diseases and tissue morphology (Wang et al., 2022c; Zhang et al., 2025; Huang et al., 2023). GLoRIA and MGCA further leverage local and multi-granularity cross-modal alignment to address the limited lesion-localization capability of image-level contrastive learning (Huang et al., 2021; Wang et al., 2022a).

More recently, VLMs have demonstrated strong capabilities in medical image understanding and clinical reasoning. LLaVA-Med improves biomedical image interpretation through vision-language alignment and instruction tuning (Li et al., 2023), while LLaVA-Rad employs a lightweight cross-modal adapter for radiological semantic modeling (Zambrano Chaves et al., 2025). These models can generate clinically informative text that provides fine-grained descriptions and lesion-specific findings for medical interpretation (Peng et al., 2023; Liu et al., 2025). Chen et al. (2026) further incorporate VLM-derived semantics with gaze-based weak supervision for medical image segmentation. However, existing VLM-based studies mainly focus on semantic understanding or gaze-assisted representation learning, rather than modeling the sequential attention allocation and transitions required for scanpath prediction.

3 METHODOLOGY

3.1 Preliminaries

Radiology scanpath prediction aims to model the visual search trajectory of a radiologist performing a specific diagnostic task. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W}$ and a target finding $c \in \mathcal{C}$, the corresponding ground-truth scanpath is represented as $\mathbf{S}^* = \{\mathbf{f}_t^*\}_{t=1}^{T}$, where $\mathbf{f}_t^* = (x_t^*, y_t^*, d_t^*)$. (x_t^*, y_t^*) denotes the 2D coordinates of the t-th fixation, d_t^* represents the corresponding fixation duration, T denotes the scanpath length, and $\mathcal{C}$ comprises 13 target finding categories. As illustrated in Fig. 1, HiSTAR generates the scanpath through step-wise decoding. At the t-th time step, conditioned on the input image, target finding, and fixation history, the model predicts the subsequent fixation heatmap $\mathbf{H}_t \in [0,1]^{H\times W}$, fixation duration $\hat{d}_t$, and termination probability $\hat{\tau}_t \in [0,1]$. For ground-truth fixation (x_t^*, y_t^*), a 2D Gaussian heatmap $\mathbf{Y}_t \in [0,1]^{H\times W}$ is constructed to provide supervision (Wang et al., 2022b).

Let Ω denote the set of valid time steps. The base prediction objective combines a pixel-wise focal loss ($\mathcal{L}_{\text{focal}}$) for fixation localization, an ℓ_1 loss for fixation duration regression, and a binary cross-entropy loss ($\mathcal{L}_{\text{BCE}}$) for scanpath termination prediction:

$$\mathcal{L}_{\text{base}} = \frac{1}{|\Omega|} \sum_{t\in\Omega} \left[\mathcal{L}_{\text{focal}}(\mathbf{H}_t, \mathbf{Y}_t) + |\hat{d}_t - d_t^*| + \mathcal{L}_{\text{BCE}}(\hat{\tau}_t, \tau_t^*) \right]. \tag{1}$$

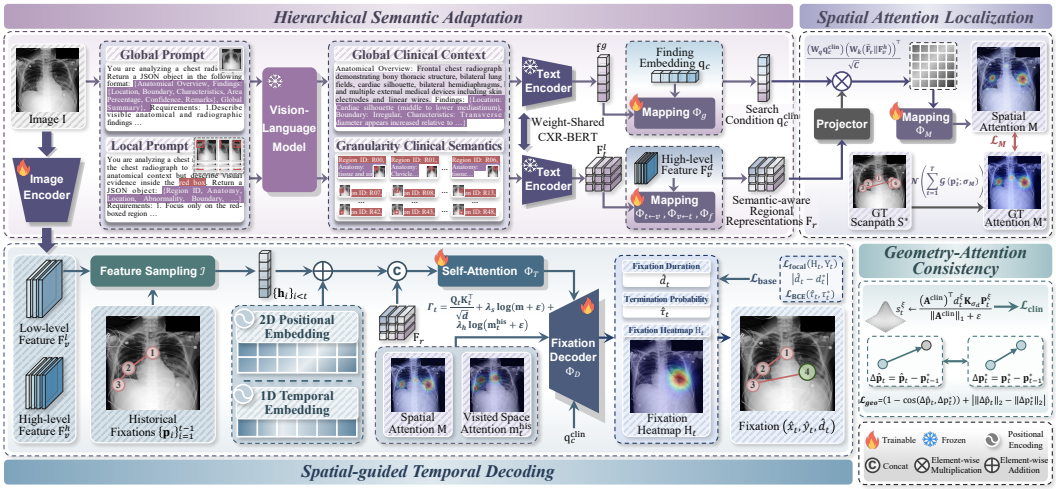


Figure 1: Overview of HiSTAR. Hierarchical semantic adaptation incorporates global context and granularity semantics to construct image-specific search conditions and semantic-aware regional representations. Spatial-guided temporal decoding estimates a task-relevant spatial prior to guide progressive trajectory generation, while geometry-attention consistency regularizes attention allocation and inter-fixation geometry.

3.2 Hierarchical Semantic Adaptation

We propose a Hierarchical Semantic Adaptation (HSA) mechanism with two complementary levels: global context adapts the target finding into an image-specific clinical search condition, while granularity semantics are integrated with visual features to form semantic-aware regional representations. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W}$, the visual backbone extracts a low-level feature $\mathbf{F}_v^l \in \mathbb{R}^{H_s\times W_s\times C}$ for attention localization and a high-level feature $\mathbf{F}_v^h \in \mathbb{R}^{H_d\times W_d\times C}$ for regional clinical conditioning. The VLM, such as Doubao-Seed-2.0, uses a global prompt to generate an anatomical overview, global summary, and finding-specific descriptions of location, boundary, characteristics, and proportion, forming the global clinical context. The complete prompt is provided in the **Appendix B**. The global context is encoded by CXR-BERT (Boecking et al., 2022) into a vector $\mathbf{f}^g \in \mathbb{R}^C$. For a target finding c, we use a learnable embedding over the 13 findings to obtain $\mathbf{q}_c \in \mathbb{R}^C$. We define a learnable mapping Φ_g, implemented with linear projection and residual connections, to calibrate the search condition:

$$\mathbf{q}_c^{\text{clin}} \leftarrow \Phi_g\left(\mathbf{q}_c, \mathbf{f}^g\right), \qquad \mathbf{q}_c^{\text{clin}} \in \mathbb{R}^C. \tag{2}$$

For localized granularity evidence, we partition the chest X-ray into a $P \times P$ grid and sequentially highlight each region with a red bounding box, with $H_d = W_d = P$. Given each marked image and a local prompt, the VLM generates region-specific descriptions of anatomy, location, abnormality, boundary, characteristics, confidence, and summary. These descriptions are encoded by CXR-BERT and spatially arranged as $\mathbf{F}_t^l \in \mathbb{R}^{P\times P\times C}$. For cross-modal interaction, visual and textual features are flattened into token sequences. The fusion process is formulated as:

$$\widetilde{\mathbf{F}}_t^l \leftarrow \Phi_{t\leftarrow v}\left(\mathbf{F}_t^l, \mathbf{F}_v^h\right), \quad \widetilde{\mathbf{F}}_v^h \leftarrow \Phi_{v\leftarrow t}\left(\mathbf{F}_v^h, \mathbf{F}_t^l\right), \quad \mathbf{F}_r \leftarrow \Phi_f\left(\widetilde{\mathbf{F}}_v^h \| \widetilde{\mathbf{F}}_t^l\right), \tag{3}$$

where $\Phi_{t\leftarrow v}$ and $\Phi_{v\leftarrow t}$ denote position-constrained cross-attention with residual normalization, and Φ_f is a dimensionality projection. The resulting $\mathbf{F}_r \in \mathbb{R}^{P\times P\times C}$ integrates local visual and clinical semantics and, together with $\mathbf{q}_c^{\text{clin}}$, guides subsequent spatial localization and trajectory generation.

3.3 Spatial-guided Temporal Decoding

We decompose scanpath prediction into spatial attention localization and temporal trajectory generation. Accordingly, Spatial-guided Temporal Decoding (STD) first identifies diagnostically relevant regions and then uses the resulting spatial attention as a prior for progressive trajectory generation.

Specifically, $\mathbf{F}_r$ is projected to the spatial resolution of $\mathbf{F}_v^l$ as $\widetilde{\mathbf{F}}_r$, and a task-specific spatial attention distribution is obtained by matching it with the clinical search condition:

$$\mathbf{M} \leftarrow \Phi_M \left(\frac{(\mathbf{W}_q \mathbf{q}_c^{\text{clin}}) \left(\mathbf{W}_k (\widetilde{\mathbf{F}}_r \| \mathbf{F}_v^l) \right)^\top}{\sqrt{C}} \right), \qquad \mathbf{M} \in [0,1]^{H \times W}, \tag{4}$$

where Φ_M, implemented with transposed convolutions, maps the responses to an input-resolution attention map. The resulting $\mathbf{M}$ is independent of fixation history and represents the overall diagnostic search space. Its ground-truth map is constructed by accumulating and normalizing Gaussian responses centered at all ground-truth fixations: $\mathbf{M}^* \leftarrow \mathcal{N}(\sum_{t=1}^{T} \mathcal{G}(\mathbf{p}_t^*; \sigma_M))$. Let $\mathbf{P}_M$ and $\mathbf{P}_M^*$ denote the normalized distributions of the predicted and ground-truth maps, respectively. The spatial attention objective is defined as:

$$\mathcal{L}_M = \frac{\|\mathbf{M} - \mathbf{M}^*\|_F^2}{HW} + D_{\text{KL}} \left(\mathbf{P}_M^* \| \mathbf{P}_M \right). \tag{5}$$

The Frobenius term constrains attention responses, while the KL divergence aligns their spatial distributions. Based on this diagnostic space, HiSTAR models sequential attention transitions. At step t, features at previous fixation locations $\mathbf{p}_i{}_{i=1}^{t-1}$ are sampled from the shallow visual feature, augmented with learnable temporal encodings, and combined with the regional representation through the hierarchical temporal mapping Φ_T:

$$\mathbf{h}_i \leftarrow \mathcal{I} \left(\mathbf{F}_v^l, \Pi_s(\mathbf{p}_i) \right) + \mathbf{e}_i, \qquad \mathbf{E}_t \leftarrow \left[\Phi_T \left(\mathbf{F}_r \| \{\mathbf{h}_i\}_{i=1}^{t-1} \right) \right]_{1:\mathrm{P}^2}, \tag{6}$$

where $\Pi_s(\cdot)$ maps fixation coordinates to the feature scale, $\mathcal{I}(\cdot)$ samples the corresponding visual features, $\mathbf{e}_i$ is the temporal encoding, and Φ_T is a self-attention-based temporal mapping. The resulting $\mathbf{E}_t$ encodes regional evidence and accumulated search history. During decoding, the spatial attention $\mathbf{M}$ and history map $\mathbf{M}_t^{\text{his}}$ represent the task-relevant diagnostic space and previously visited regions, respectively. They are projected to the decoder resolution as $\mathbf{m}$ and $\mathbf{m}_t^{\text{his}}$, and incorporated as spatial biases into the attention logits at step t:

$$\mathbf{\Gamma}_t = \frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d}} + \lambda_s \log(\mathbf{m} + \varepsilon) + \lambda_h \log(\mathbf{m}_t^{\text{his}} + \varepsilon), \qquad \mathbf{z}_t \leftarrow \Phi_D \left(\mathbf{\Gamma}_t, \mathbf{V}_t \right), \tag{7}$$

where Φ_D normalizes $\mathbf{\Gamma}_t$ and aggregates the value representations. The coefficients $\lambda_s > 0$ and $\lambda_h < 0$, constrained through Softplus parameterization, respectively emphasize task-relevant regions and suppress heavily visited regions. Finally, $\mathbf{z}_t$ is passed through separate heads to predict the fixation heatmap $\mathbf{H}_t$, duration $\hat{d}_t$, and termination probability $\hat{\tau}_t$ for progressive scanpath generation.

3.4 Geometry-Attention Consistency

Existing methods mainly focus on fixation localization, with limited modeling of attention allocation and transition geometry. To address this, we propose Geometry-Attention Consistency (GAC), which constrains scanpaths through task-relevant attention allocation and saccadic geometry. At time step t, the predicted fixation heatmap $\mathbf{H}_t$, ground-truth heatmap $\mathbf{Y}_t$, and spatial attention map $\mathbf{M}$ from STD are mapped to a common spatial resolution:

$$\left(\hat{\mathbf{P}}_t, \mathbf{P}_t^*, \mathbf{A}^{\text{clin}} \right) \leftarrow \mathcal{R}_P \left(\mathbf{H}_t, \mathbf{Y}_t, \text{sg}[\mathbf{M}] \right), \tag{8}$$

where $\mathcal{R}_P$ denotes spatial downsampling to a $P \times P$ resolution, and $\text{sg}[\cdot]$ denotes the stop-gradient operation. To encode spatial proximity among regions, we construct a Gaussian kernel as $[\mathbf{K}_{\sigma_d}]_{ij} \leftarrow \exp(-\|\mathbf{r}_i - \mathbf{r}_j\|_2^2 / 2\sigma_d^2)$, where σ_d controls the spatial propagation range. Here, $N = P^2$ is the number of regions, and $\mathbf{r}_i = (x_i^r, y_i^r) \in [0,1]^2$ denotes the normalized center of region i. The attention fields and the attention values for the predicted and ground-truth are formulated as:

$$\mathbf{D}_t^\xi \leftarrow d_t^\xi \, \mathbf{K}_{\sigma_d} \mathbf{P}_t^\xi, \qquad s_t^\xi \leftarrow \frac{(\mathbf{A}^{\text{clin}})^\top \mathbf{D}_t^\xi}{\|\mathbf{A}^{\text{clin}}\|_1 + \varepsilon}, \qquad \xi \in \{\hat{}, ^*\}. \tag{9}$$

By combining spatial allocation, fixation duration, and task relevance, s_t measures attention assigned to diagnostically relevant regions. An ℓ_1 distance between $\hat{s}_t$ and s_t^* defines the attention

consistency loss $\mathcal{L}_{\text{clin}}$. To model inter-fixation transitions, let $\bar{\mathbf{H}}_t$ be the normalized predicted fixation distribution and $\mathbf{z}xy \in [0,1]^2$ the normalized pixel coordinate. The expected fixation is $\hat{\mathbf{p}}_t = \sum_{x,y} \bar{\mathbf{H}}_t(x,y)\mathbf{z}_{xy}$. Using $\mathbf{p}^*_{t-1}$ as a common anchor, we define $\Delta\hat{\mathbf{p}}_t = \hat{\mathbf{p}}_t - \mathbf{p}^*_{t-1}$ and $\Delta\mathbf{p}^*_t = \mathbf{p}^*_t - \mathbf{p}^*_{t-1}$. The geometric consistency loss is defined as:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\Omega_{\text{geo}}|} \sum_{t \in \Omega_{\text{geo}}} \left[1 - \cos\left(\Delta\hat{\mathbf{p}}_t, \Delta\mathbf{p}^*_t\right) + \left|\|\Delta\hat{\mathbf{p}}_t\|_2 - \|\Delta\mathbf{p}^*_t\|_2\right|\right], \tag{10}$$

where Ω_{geo} denotes the valid transition steps. The cosine and amplitude terms constrain saccade direction and displacement, respectively. The GAC objective is defined as:

$$\mathcal{L}_{\text{GAC}} = \mathcal{L}_{\text{clin}} + \mathcal{L}_{\text{geo}}. \tag{11}$$

The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_M \mathcal{L}_M + \lambda_G \mathcal{L}_{\text{GAC}}. \tag{12}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Scanpath prediction experiments are conducted on GazeSearch (Pham et al., 2025c), which is curated from EGD (Karargyris et al., 2021; Johnson et al., 2019) and REFLACX (Lanfredi et al., 2022) and contains chest X-rays with task-relevant fixation sequences across multiple finding categories. We follow the official training, validation, and test splits. To evaluate the diagnostic value of predicted scanpaths, we conduct chest X-ray classification on SIIM-ACR and the self-collected TB-MOUSE dataset. For cross-domain evaluation, we further use COCO-Search18 (Chen et al., 2021b), a natural-image visual search benchmark. See the **Appendix A.1** for detailed dataset descriptions.

Evaluation Metrics. Following the GazeSearch protocol, we adopt ScanMatch (Cristino et al., 2010; Sogo, 2013), MultiMatch (Dewhurst et al., 2012), String-Edit Distance (SED) (Brandt & Stark, 1997; Foulsham & Underwood, 2008), and Scaled Time-Delay Embedding (STDE) (Wang et al., 2011) to evaluate scanpaths. ScanMatch measures similarity using the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) and is reported under two settings: without duration (w/o D.) and with duration (w/ D.). MultiMatch evaluates similarity across five continuous dimensions: vector, direction, length, position, and duration. For downstream classification, we report Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and F1 Score (F1).

Implementation Details. All images are resized to 224×224 pixels. An ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) is adopted as the visual backbone. Clinical descriptions are generated offline using Doubao-Seed-2.0 for its multimodal understanding and structured instruction-following capabilities, and encoded using CXR-BERT (Boecking et al., 2022). Granularity clinical semantics are organized according to a 7×7 grid. The model is trained for 50 epochs using the AdamW optimizer, with the initial learning rate and weight decay set to 1×10^{-4}. All experiments are implemented in PyTorch and conducted on an NVIDIA GeForce RTX 4090 GPU. The loss weights λ_M and λ_G are both set to 0.5.

4.2 COMPARISON WITH STATE-OF-THE-ART SCANPATH PREDICTION METHODS

Quantitative Results. We compare HiSTAR against IRL (Yang et al., 2020), FFMs (Yang et al., 2022), ChenLSTM (Chen et al., 2021a), ChenLSTM-ISP (Chen et al., 2024), Gazeformer (Mondal et al., 2023), Gazeformer-ISP (Chen et al., 2024), HAT (Yang et al., 2024), LogitGaze (Lvov & Pershin, 2025a), ChestSearch (Pham et al., 2025c), and LogitGaze-Med (Lvov & Pershin, 2025b). For LogitGaze-Med, we report the results obtained from our reproduction under the same protocol. As reported in Table 1, HiSTAR achieves the best performance, with significant improvements in ScanMatch, SED, and STDE confirmed by the Wilcoxon signed-rank test ($p < 0.05$). Under the w/o D. and w/ D. settings, HiSTAR obtains ScanMatch scores of 0.3942 and 0.3664, exceeding ChestSearch by 0.0621 and 0.1432, respectively. HiSTAR achieves consistent improvements across all MultiMatch dimensions, with particularly strong gains in position and duration, indicating more accurate spatial localization and fixation-duration modeling.

Table 1: Quantitative results on the GazeSearch dataset. Bold indicates the best results.

Method	ScanMatch↑		MultiMatch↑					SED↓	STDE↑
	w/o D.	w/ D.	Vector	Direction	Length	Position	Duration		
IRL	0.1495	-	0.8248	0.6402	0.7688	0.6998	-	6.6250	0.7043
FFMs	0.2766	-	0.8914	0.6576	0.8785	0.8140	-	5.9221	0.8055
ChenLSTM	0.2751	0.2153	0.8825	0.6222	0.8731	0.7940	0.6384	5.3468	0.7841
ChenLSTM-ISP	0.2863	0.2205	0.8847	0.6430	0.8721	0.7980	0.6504	5.2895	0.7865
Gazeformer	0.2971	0.2042	0.9080	0.6506	0.9035	0.8147	0.5901	5.1024	0.8030
Gazeformer-ISP	0.2736	0.2146	0.9038	0.6181	0.8892	0.8031	0.6755	5.1905	0.7875
HAT	0.3120	-	0.9064	0.6443	0.9065	0.8138	-	5.0613	0.8006
LogitGaze	0.3278	0.2247	0.8845	0.6465	0.9137	0.8124	0.6272	5.1081	0.8069
ChestSearch	0.3321	0.2232	0.9173	0.6790	0.9174	0.8293	0.6951	4.8831	0.8089
LogitGaze-Med	0.3572	0.2760	0.9092	0.6764	0.9153	0.8362	0.7163	4.9147	0.8124
Ours	**0.3942**	**0.3664**	**0.9187**	**0.6828**	**0.9194**	**0.8546**	**0.7883**	**4.8709**	**0.8375**

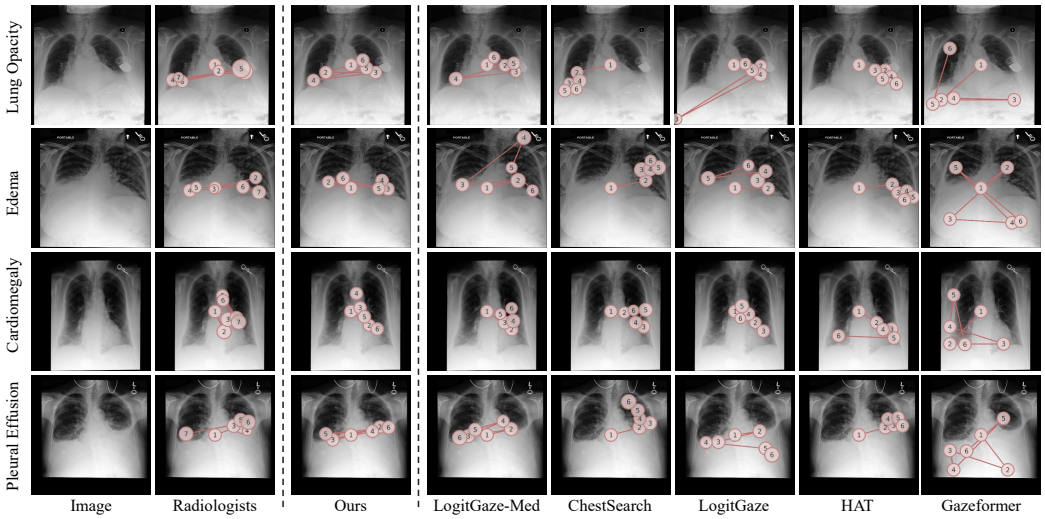


Figure 2: Qualitative comparison of predicted scanpaths on the GazeSearch dataset.

Qualitative Results. Fig. 2 presents qualitative comparisons. HiSTAR concentrates fixations on diagnostically relevant regions and produces transitions that more closely follow the ground-truth scanpaths. In contrast, competing methods often exhibit long saccades across background areas or deviate from the target. For findings such as pulmonary edema, HiSTAR provides more compact spatial coverage and smoother inter-fixation transitions, better reflecting the diagnostic search process. Fig. 3 further demonstrates that the spatial attention maps predicted by HiSTAR reproduce the major high-response regions observed in the ground-truth attention maps, with comparable peak locations and overall spatial patterns. The qualitative results suggest that the trajectory-level spatial prior learned by STD provides a stable estimate of diagnostic attention, thereby reducing trajectory prediction errors.

4.3 Chest X-ray Classification Analysis

Table 2 evaluates scanpaths from different sources on SIIM-ACR and TB-MOUSE, where CS. denotes ChestSearch. We select GA-Net (Wang et al., 2022b) and EG-ViT (Ma et al., 2023) as representative CNN- and Transformer-based gaze-guided classifiers, respectively, to evaluate whether HiSTAR-predicted gaze remains effective across different model architectures. Predicted scanpaths are converted into attention maps and provided to the gaze-guided classifiers GA-Net (Wang et al., 2022b) and EG-ViT (Ma et al., 2023). On SIIM-ACR, HiSTAR-predicted gaze improves accuracy from 83.20% to 84.00% in GA-Net. On TB-MOUSE, it raises accuracy to 89.84% and F1 to 89.82%, outperforming ChestSearch-based gaze by 1.41% and 1.49%, respectively. Similar gains are observed with EG-ViT. Although human attention annotations remain the upper bound, HiSTAR-predicted gaze can serve as a useful auxiliary attention prior for disease classification.

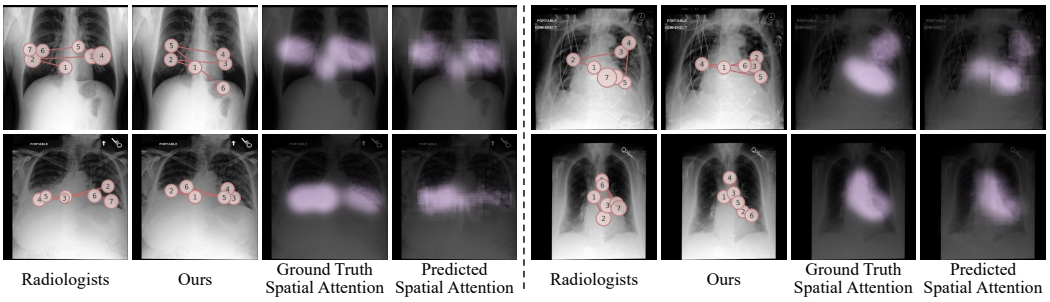


Figure 3: Qualitative comparison of predicted and ground-truth spatial attention on the GazeSearch dataset.

Table 2: Performance comparison on SIIM-ACR and TB-MOUSE datasets.

	Acc↑	AUC↑	F1↑
SIIM-ACR			
ResNet	83.20	71.42	81.08
GA-Net (CS.)	83.60	71.25	83.29
GA-Net (Ours)	84.00	71.53	83.63
GA-Net (Gaze)	84.80	72.68	84.03
ViT	83.60	74.16	83.77
EG-ViT (CS.)	84.00	73.26	83.71
EG-ViT (Ours)	84.80	74.89	84.03
EG-ViT (Gaze)	85.60	75.30	85.14
TB-MOUSE			
ResNet	87.86	96.06	87.80
GA-Net (CS.)	88.43	95.65	88.33
GA-Net (Ours)	89.84	96.57	89.82
GA-Net (Gaze)	90.59	96.99	90.48
ViT	88.23	95.64	88.20
EG-ViT (CS.)	88.45	95.77	88.53
EG-ViT (Ours)	89.18	96.42	89.05
EG-ViT (Gaze)	89.55	96.61	89.56

Table 3: Ablation study of different loss components in HiSTAR. MM. denotes the average MultiMatch score.

$\mathcal{L}_{\text{base}}$	$\mathcal{L}_M$	$\mathcal{L}_{\text{geo}}$	$\mathcal{L}_{\text{clin}}$	ScanMatch↑ w/o D.	ScanMatch↑ w/ D.	MM.↑	SED↓	STDE↑
✓				0.3614	0.3027	0.8125	5.0842	0.8116
✓	✓			0.3823	0.3415	0.8227	4.9436	0.8245
✓	✓	✓		0.3886	0.3541	0.8254	4.8917	0.8314
✓	✓		✓	0.3867	0.3472	0.8279	4.9073	0.8298
✓	✓	✓	✓	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**

Table 4: Ablation study of global context and granularity semantics and the guided attention mechanism.

Global	Gran.	Γ_t	ScanMatch↑ w/o D.	ScanMatch↑ w/ D.	MM.↑	SED↓	STDE↑
✗	✗	**w/ m.**	0.3673	0.3360	0.8203	5.1623	0.8193
✓	✗	**w/ m.**	0.3735	0.3447	0.8243	4.9862	0.8246
✗	✓	**w/ m.**	0.3856	0.3518	0.8304	4.9157	0.8322
✓	✓	**w/o m.**	0.3841	0.3536	0.8267	4.9124	0.8279
✓	✓	**w/ m.**	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**

Table 5: Cross-domain evaluation on COCO-Search18 dataset with mean ± standard deviation.

Method	Training Domain	ScanMatch w/o D.↑	ScanMatch w/ D.↑	MultiMatch Avg↑
Gazeformer	COCO-Search18	0.492 ± 0.014	0.441 ± 0.015	0.816 ± 0.010
LogitGaze-Med	GazeSearch	0.353 ± 0.018	0.289 ± 0.020	0.749 ± 0.014
HiSTAR	GazeSearch	0.371 ± 0.016	0.312 ± 0.018	0.760 ± 0.016

4.4 Ablation Analysis

Ablation on Loss Functions. Table 3 reports ablations of the loss terms on the GazeSearch dataset, where MM. denotes the average MultiMatch score. Starting from the base loss with an MM. score of 0.8125, adding the spatial attention loss $\mathcal{L}_M$ improves the score to 0.8227, indicating better spatial localization. Further incorporating the geometric loss raises MM. to 0.8254 by constraining saccade direction and amplitude. The full model achieves the best score of 0.8328, demonstrating the complementary benefit of attention and geometric consistency.

Ablation on HSA and Spatial Guidance. Table 4 evaluates global context, granularity semantics, and spatial-prior guidance on the GazeSearch dataset. Granularity semantics provide stronger localized anatomical and abnormality-related cues, while combining both semantic levels further improves performance. Introducing spatial guidance yields consistent gains across all metrics. These results demonstrate that hierarchical semantics provide complementary diagnostic evidence, while spatial guidance effectively supports subsequent trajectory generation. Fig. 4 presents representative

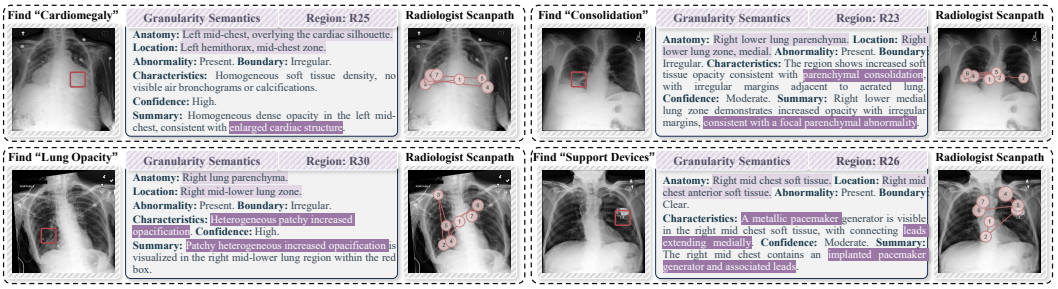
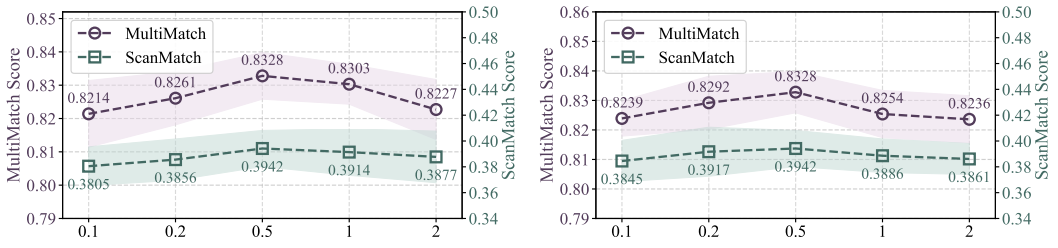


Figure 4: Examples of granularity semantics generated by the VLM. Regional descriptions capture anatomical and abnormality-related information.



(a) λ_M ($\lambda_G = 0.5$) (b) λ_G ($\lambda_M = 0.5$)

Figure 5: Sensitivity analysis of the loss weights λ_M and λ_G. ScanMatch w/o D. and average MultiMatch are reported by varying the loss weights.

granularity-level clinical semantics, including findings such as cardiac enlargement, consolidation, opacity, and pacemaker presence. These descriptions are spatially aligned with both the target finding and radiologist-inspected regions, providing qualitative evidence that HSA derives image- and region-specific semantic cues for subsequent STD.

Loss Weight Analysis. We evaluate the sensitivity to λ_M and λ_G over $\{0.1, 0.2, 0.5, 1.0, 2.0\}$ using ScanMatch w/o D. and average MultiMatch. As shown in Fig. 5, both settings achieve the best performance at 0.5, with only modest variations across the tested range. These results indicate that HiSTAR is relatively robust to the choice of loss weights.

Cross-Domain Generalization Analysis. Table 5 evaluates cross-domain generalization from chest X-rays to natural scenes. The standard deviation is computed across individual samples. GazeFormer is trained on COCO-Search18, whereas LogitGaze-Med and HiSTAR are trained only on GazeSearch. HiSTAR achieves ScanMatch scores of 0.371 and 0.312, outperforming LogitGaze-Med. The advantage suggests that granularity semantics enhance regional representations with task-related cues, while STD improves transferable visual search modeling by using spatial priors to guide temporal trajectory prediction.

5 Conclusion & Limitations

We propose HiSTAR, a Hierarchical Semantic-aware Spatio-Temporal Decoding framework for radiology scanpath prediction. HiSTAR leverages VLMs to incorporate global clinical semantics and fine-grained regional information. Hierarchical Semantic Adaptation enhances task-relevant representations, while Spatial-guided Temporal Decoding first identifies diagnostically relevant spatial regions and then progressively generates the visual search trajectory. Experiments on GazeSearch show that HiSTAR outperforms existing methods across multiple scanpath similarity metrics. Downstream classification experiments on SIIM-ACR and TB-MOUSE further validate the effectiveness of the predicted gaze information for disease recognition.

HiSTAR relies on VLMs to generate clinical semantics, which introduces additional computational overhead. Moreover, due to data limitations, the current evaluation mainly focuses on chest X-ray images, and cross-modality generalization remains to be further investigated. Future work will explore more efficient and reliable clinical semantic extraction strategies.

AI USE STATEMENT

In this work, we used generative AI tools for translation assistance. We have not used generative AI tools for generating synthetic datasets, developing theoretical frameworks, proposing hypotheses, designing experiments or interpreting results, and other remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for drafting partial manuscript text, editing the paper for readability. We have reviewed all AI-assisted work. All AI-polished text was manually revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments were conducted strictly in accordance with relevant ethical guidelines and institutional review processes. The TB-MOUSE data were collected from a radiologist participating in this study. The institutional ethics committee reviewed and approved the study protocol, and the participant provided informed consent. All datasets employed were obtained and used in accordance with relevant usage guidelines. No personally identifiable information is disclosed or utilized. During the research process, every effort was made to mitigate potential biases and discriminatory outcomes, and no experiments posing security risks were conducted.

REPRODUCIBILITY STATEMENT

We have made every reasonable effort to support the reproducibility of the experimental results reported in this paper. Our code is publicly available in an anonymous repository to facilitate reproduction. The experimental setup, including training configurations and parameter settings, is elaborated in the main text. We also provide a comprehensive description of HiSTAR to help other researchers replicate our experiments.

REFERENCES

Susmit Agrawal, Matthias Bethge, and Matthias Kümmerer. DeepGaze3.5-VL: Modeling scanpaths via autoregressive token prediction. In *Computer Vision – ECCV 2026*, pp. 242–269, Cham, 2026. Springer Nature Switzerland.

Akash Awasthi, Ngan Le, Zhigang Deng, Carol C. Wu, and Hien Van Nguyen. Collaborative integration of AI and human expertise to improve detection of chest radiograph abnormalities. *Radiology: Artificial Intelligence*, 7(5):e240277, 2025.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *Computer Vision – ECCV 2022*, volume 13696 of *Lecture Notes in Computer Science*, pp. 1–21, Cham, 2022. Springer Nature Switzerland.

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.

Jingkun Chen, Haoran Duan, Xiao Zhang, Boyan Gao, Vicente Grau, and Jungong Han. From gaze to insight: Bridging human visual attention and vision language model explanation for weakly-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(3):1051–1061, 2026.

Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10871–10880, 2021a.

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 (CVPR)*, pp. 25420–25431, June 2024.

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):8776, 2021b.

Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D. Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3):692–700, 2010.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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(4):1079–1100, 2012.

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, 2008.

Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):61–84, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931, 2021.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, A de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

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):92, 2021.

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, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

Yi Liu, Chengxin Li, Shoukun Xu, and Jungong Han. Part-whole relational fusion towards multi-modal scene understanding. *International Journal of Computer Vision*, 133(7):4483–4503, 2025.

Jianxun Lou, Huasheng Wang, Xinbo Wu, John Cho Hui Ng, Richard White, Kaveri A. Thakoor, Padraig Corcoran, Ying Chen, and Hantao Liu. Chest X-ray visual saliency modeling: Eye-tracking dataset and saliency prediction model. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16920–16930, 2025.

Dmitry Lvov and Ilya Pershin. LogitGaze: Predicting human attention using semantic information from vision-language models. In *Workshop on Reasoning and Planning for Large Language Models at ICLR*, 2025a.

Dmitry Lvov and Ilya Pershin. From human attention to diagnosis: Semantic patch-level integration of vision-language models in medical imaging. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 131558–131579. Curran Associates, Inc., 2025b.

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):3384–3394, 2023.

Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1441–1450, 2023.

Sounak Mondal, Naveen Sendhilnathan, Ting Zhang, Yue Liu, Michael Proulx, Michael Louis Iuzzolino, Chuan Qin, and Tanya R. Jonker. Gaze-language alignment for zero-shot prediction of visual search targets from human gaze scanpaths. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2738–2749, October 2025.

Shravan Murlidaran and Miguel P Eckstein. Eye movements during free viewing to maximize scene understanding. *Nature Communications*, 17(1):940, 2025.

T Nan, S Zheng, S Qiao, H Quan, X Gao, J Niu, B Zheng, C Guo, Y Zhang, X Wang, L Zhao, Z Wu, Y Guo, X Li, M Zou, S Ning, Y Zhao, W Qian, H Chen, R Qi, X Gao, and X Cui. Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. *Nature Communications*, 16(1):5493, 2025.

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

Luke Palmer, Petar Palasek, and Hazem Abdelkawy. Beyond scanpaths: Graph-based gaze simulation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15199–15211, June 2026.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1):210, 2023.

Trong Thang Pham, Akash Awasthi, Saba Khan, Esteban Duran Marti, Tien-Phat Nguyen, Khoa Vo, Minh Tran, Son Nguyen, Cuong Tran, Yuki Ikebe, Anh Totti Nguyen, Anh Nguyen, Zhigang Deng, Carol C. Wu, Hien Nguyen, and Ngan Le. CT-ScanGaze: A dataset and baselines for 3D volumetric scanpath modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21732–21743, 2025a.

Trong-Thang Pham, Anh Nguyen, Zhigang Deng, Carol C Wu, Hien Van Nguyen, and Ngan Le. Interpreting radiologist's intention from eye movements in chest X-ray diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 257–266. ACM, 2025b.

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 *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 96–106, February 2025c.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Hiroyuki Sogo. GazeParser: An open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Research Methods*, 45(3):684–695, 2013.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33536–33549. Curran Associates, Inc., 2022a.

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, 2022b.

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*, pp. 441–448, 2011.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics.

Shaoxuan Wu, Xiao Zhang, Jingkun Chen, Xinyu Zhang, Yaqiong Xing, and Jun Feng. GiTNet: A graph-based trajectory-informed network for gaze-supervised medical image segmentation. *Medical Image Analysis*, 112:104112, 2026a. ISSN 1361-8415.

Shaoxuan Wu, Xiao Zhang, Xiaodi Zhao, Yunzhi Tian, Yilin Tang, and Jun Feng. From spatial semantics to temporal context: Leveraging gaze trajectory for weakly supervised medical image segmentation, 2026b.

Ruoyu Xue, Jingyi Xu, Sounak Mondal, Hieu Le, Greg Zelinsky, Minh Hoai, and Dimitris Samaras. Few-shot personalized scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13497–13507, June 2025.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 193–202, 2020.

Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *Computer Vision – ECCV 2022*, volume 13664 of *Lecture Notes in Computer Science*, pp. 52–68. Springer, 2022.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1683–1693, 2024.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications*, 16(1):3108, 2025.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. SIIM-ACR pneumothorax segmentation, 2019. URL https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco,

Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025.

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. MedEyes: Learning dynamic visual focus for medical progressive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13916–13924, 2026.

APPENDIX

A EXPERIMENT DESCRIPTION

A.1 DATASETS

GazeSearch. GazeSearch (Pham et al., 2025c) is a benchmark for visual search in chest X-ray diagnosis, constructed from two eye-tracking datasets, EGD (Karargyris et al., 2021) and REFLACX (Lanfredi et al., 2022). Specifically, EGD contributes 413 images and REFLACX contributes 1,668 images. The dataset maps free-viewing recordings into visual search sequences aligned with specified findings and constrains the scanpaths based on observations that later fixations tend to approach discriminative regions and that fixations within lesions generally have longer durations. Each sample contains 1-6 findings, and the maximum sequence length is set to 7, including a default initial fixation at the image center. In total, the dataset contains 4,875 scanpath records with 29,118 fixations. We follow the official split, consisting of 1,456 training samples (70%), 208 validation samples (10%), and 417 test samples (20%).

The label taxonomy of GazeSearch is derived from the 14 standard CheXpert findings. Since GazeSearch adopts a target-present visual search setting, **[No Finding]** is not used as a search target, resulting in 13 categories for task conditioning. The complete set of 14 labels includes **[Atelectasis]**, **[Cardiomegaly]**, **[Consolidation]**, **[Edema]**, **[Enlarged Cardiomediastinum]**, **[Fracture]**, **[Lung Lesion]**, **[Lung Opacity]**, **[Pleural Effusion]**, **[Pleural Other]**, **[Pneumonia]**, **[Pneumothorax]**, **[Support Devices]**, and **[No Finding]**.

SIIM-ACR. SIIM-ACR (Zawacki et al., 2019) contains 1,250 chest X-ray images with gaze annotations, categorized as normal or pneumothorax. We follow the official split of 1,000 training images and 250 test images. The dataset uses real eye-tracking fixations as visual attention signals.

TB-MOUSE. TB-MOUSE contains 1,000 training images and 2,200 test images, each labeled as either tuberculosis or normal. Mouse trajectories are available for all 1,000 training images and were collected from an experienced radiologist during routine chest X-ray interpretation. Compared with dedicated eye-tracking systems, mouse-based interaction provides a more accessible and readily deployable means of capturing visual search behavior. During data acquisition, the system continuously records cursor positions while the radiologist inspects and interacts with each chest radiograph.

To approximate fixation behavior from mouse trajectories, stationary points are first identified from the continuous cursor sequence using the I2MC algorithm. A two-dimensional Gaussian kernel with a radius of 150 pixels and a standard deviation of 25 pixels is subsequently centered at each detected stationary point to construct a continuous spatial attention map. The resulting representation leverages eye-hand coordination by using cursor dwell patterns and movements as proxies for explicit visual search behavior. TB-MOUSE will be publicly released upon acceptance of the paper.

Using ground-truth class labels to determine target queries would allow generated scanpaths to encode category information explicitly, thereby introducing label leakage into downstream classification. To avoid such leakage, a fixed search target is assigned to all images within each dataset: **[Pneumothorax]** for SIIM-ACR and **[Lung Opacity]** for TB-MOUSE. Because the tuberculosis category is not observed during training on GazeSearch, inference on TB-MOUSE reuses the learned **[Lung Opacity]** task query $\mathbf{q}_c$. Active pulmonary tuberculosis commonly manifests on chest radiographs as parenchymal opacities, including infiltrative, exudative, consolidative, or nodular increases in density. As a broad radiographic finding encompassing these manifestations, Lung

Opacity provides a clinically reasonable search proxy for localizing potentially tuberculosis-related abnormalities.

COCO-Search18. COCO-Search18 (Chen et al., 2021b) is a target-driven visual search dataset for natural scenes, comprising 10 participants who searched for 18 target object categories across 6,202 COCO images. In our study, we use COCO-Search18 exclusively to evaluate cross-domain transfer.

For semantic generation, we adapt the prompts to the natural-image domain while preserving the overall HiSTAR framework. The chest X-ray prompts contain domain-specific medical constraints and are therefore replaced with natural-scene counterparts. The Global Prompt is reformulated to describe the overall scene layout, visible objects, and other scene-level characteristics. The Local Prompt retains the red bounding boxes used to specify individual regions in the 7×7 grid, but instead requests descriptions of object categories, appearance, and other visual attributes within each highlighted region. All remaining components of the HiSTAR architecture are kept unchanged.

A.2 Metric Details

ScanMatch. ScanMatch (Cristino et al., 2010; Sogo, 2013) discretizes continuous fixation coordinates into a 14×14 spatial grid and performs global sequence alignment using Needleman-Wunsch (Needleman & Wunsch, 1970) dynamic programming. The resulting score is normalized by the maximum value of the substitution matrix and the length of the longer sequence, where a higher value indicates greater similarity between the predicted and ground-truth scanpaths. We report ScanMatch under two evaluation settings: without fixation duration (w/o D.) and with fixation duration (w/ D.).

String-Edit Distance. String-Edit Distance (SED) (Brandt & Stark, 1997; Foulsham & Underwood, 2008) first discretizes each scanpath into a sequence of spatial symbols and then computes the number of insertion, deletion, and substitution operations required to transform the ground-truth sequence into the predicted sequence. Our implementation reports the raw Levenshtein distance rather than a normalized variant. Consequently, a lower SED indicates greater similarity between the two scanpaths.

Scaled Time-Delay Embedding. Scaled Time-Delay Embedding (STDE) (Wang et al., 2011) evaluates trajectory similarity by constructing consecutive subsequences over multiple sliding-window lengths. For each window in the second scanpath, the method identifies the window in the first scanpath that minimizes the mean point-wise spatial distance. The resulting distance is converted into a similarity score through an exponential kernel, and the scores are subsequently averaged across window lengths. STDE therefore captures both local spatial structure and sequential search dynamics, with a higher value indicating closer agreement between the predicted trajectory and the radiologist's dynamic visual search pattern.

MultiMatch. MultiMatch (Dewhurst et al., 2012) compares two scanpaths in continuous space across five complementary dimensions: vector shape (Vector / Shape), saccade direction (Direction), saccade length (Length), absolute fixation position (Position), and fixation duration (Duration). In our implementation, saccade vectors are first monotonically aligned through minimum-cost matching. The median differences in vector shape, angle, length, position, and relative fixation duration are then computed separately and normalized to derive the corresponding similarity scores. Higher values indicate greater similarity for all five dimensions.

Downstream Classification Metrics. For downstream classification on SIIM-ACR and TB-MOUSE, performance is evaluated using Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and F1 Score (F1).

B Hierarchical Semantic Acquisition with VLMs

B.1 Global Context

The Global Context module aims to generate a holistic description of the chest X-ray. The VLM output is constrained to a structured JSON format containing `[anatomical overview]`, `[findings]`, and `[global summary]`. Each finding further contains `[location]`,

`[boundary]`, `[characteristics]`, `[area percentage]`, `[confidence]`, and `[remarks]`. The complete **`[GLOBAL PROMPT]`** is provided below.

Global Prompt

```
You are analyzing a chest radiograph for representation learning.
Describe only image−observable clinical information.
Return ONLY a JSON object in the following format:

{
 anatomical overview: ...,
 findings: [
  {
   location: ...,
   boundary: clear | irregular | ambiguous | not applicable,
   characteristics: ...,
   area percentage: ...,
   confidence: high | moderate | low,
   remarks: ...
  }
 ],
 global summary: ...
}

Requirements:
1. Describe visible anatomical and radiographic findings only.
2. For each finding, explicitly describe its location and visual appearance.
3. Do not provide differential diagnoses or treatment recommendations.
4. If a characteristic cannot be reliably determined, use ‘‘uncertain’’.
5. Do not invent abnormalities that are not visually supported.
```

Figure 6: Global Prompt used to generate global clinical context from chest radiographs.

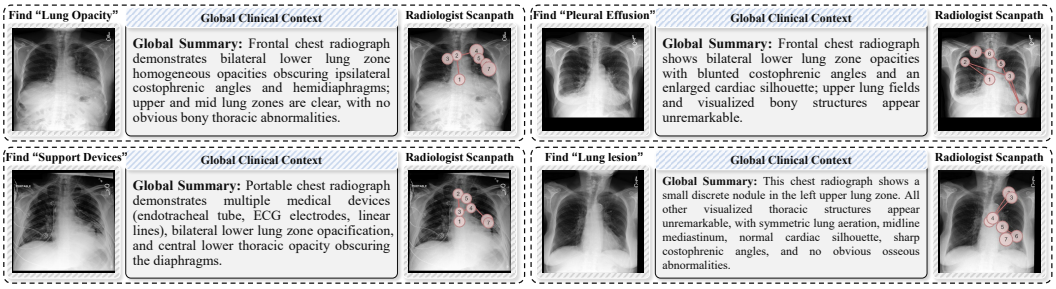


Figure 7: Representative examples in which the generated global summary is consistent with the corresponding radiographic findings.

In the first case of Fig. 7, the global context captures homogeneous high-density opacities in the bilateral lower lung fields, accompanied by obscuration of the costophrenic angles and diaphragmatic contours, while correctly indicating relatively clear middle and upper lung fields. The generated description therefore provides informative global distribution cues for the *Lung Opacity* task query. In the second case, the VLM simultaneously identifies bilateral lower-lung abnormalities, blunting of the costophrenic angles, and enlargement of the cardiac silhouette, demonstrating the capacity of global context to characterize multiple coexisting radiographic findings. In the third case, the VLM recognizes several support devices, including an endotracheal tube, electrocardiographic electrodes, and linear catheters, while retaining clinically relevant information regarding high-density changes in the bilateral lower lung fields. The fourth case presents a small, well-circumscribed nodule in the left upper lung. The global summary not only identifies the focal abnormality but also preserves the overall status of the remaining lung fields, cardiac silhouette, mediastinum, costophrenic angles, and osseous structures, thereby providing a more fine-grained global context for the *Lung Lesion* query.

B.2 GRANULARITY CLINICAL SEMANTICS

Granularity clinical semantics establishes a one-to-one correspondence between local image regions and their associated clinical textual evidence. Following the predefined single-region template, the VLM is instructed to analyze only the observable evidence within the red bounding box, while the remainder of the image serves exclusively as anatomical context for spatial localization. The corresponding **[LOCAL PROMPT]** is provided below.

Local Prompt

You are analyzing a chest radiograph.
The RED bounding box marks the ONLY local region that should be analyzed.
Use the full chest radiograph only to understand the anatomical context, but describe visual evidence inside the red box only.
Return ONLY the following JSON:

{
region id: {region id},
anatomy: ...,
location: ...,
abnormality: present | absent | uncertain,
boundary: clear | irregular | ambiguous | not applicable,
characteristics: ...,
confidence: high | moderate | low,
summary: ...
}

Requirements:
1. Focus only on the red−boxed region.
2. Use the rest of the image only for anatomical orientation.
3. Describe image−observable findings, not patient history.
4. Do not describe abnormalities outside the box.
5. If the local evidence is insufficient, return "uncertain".

Figure 8: Local prompt used to generate granularity clinical semantics from marked chest X-ray regions.

For local semantic generation and encoding, each chest radiograph is first converted to RGB and uniformly resized to 512×512 for semantic generation, whereas the corresponding model input is resized to 224×224 during training. The image is uniformly partitioned into a 7×7 grid following row-major order, yielding 49 regions indexed from R00 to R48. Each region is individually highlighted with a red bounding box and processed by the VLM to generate structured descriptions containing `[anatomy]`, `[location]`, `[abnormality]`, `[boundary]`, `[characteristics]`, `[confidence]`, and `[summary]`. A frozen BiomedVLP-CXR-BERT-specialized encoder subsequently maps both global and local textual descriptions into 128-dimensional embeddings. The 7×7 granularity is selected to match the spatial resolution of the visual feature map produced by `layer4` of ResNet-50, establishing direct spatial correspondence between the i-th semantic token and the i-th visual token.

As shown in Fig. 9, the four examples cover pulmonary parenchymal opacity, costophrenic-angle changes associated with pleural effusion, support devices, and focal lung lesions, respectively. For R26, the VLM identifies a homogeneous high-density opacity in the left lower lung with high confidence. For R29, mild hazy opacity in the right lower lung is associated with blurring of the costophrenic recess margin. For R40, the VLM recognizes a linear high-density catheter structure. In the fourth case, R18 localizes a small, well-circumscribed nodule in the left suprahilar region with high confidence. Compared with whole-image summaries, such region-level descriptions preserve more precise information about anatomical location, lesion boundaries, and morphological characteristics, thereby providing localized and interpretable semantic cues for the subsequent position-constrained bidirectional cross-modal attention.

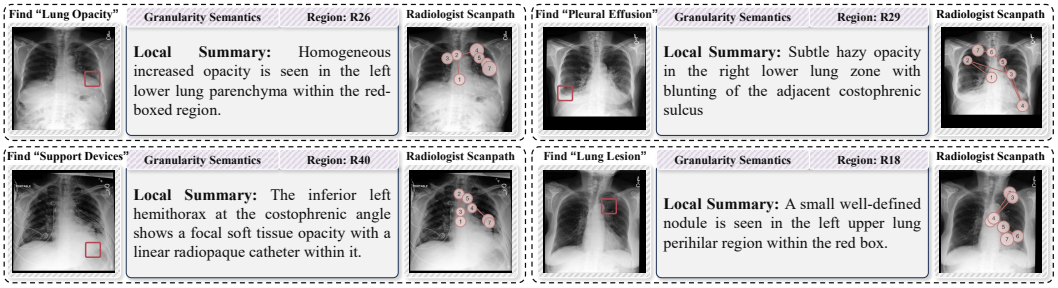


Figure 9: Representative examples of local clinical semantics on the GazeSearch dataset, where the red bounding box indicates the corresponding region in the 7×7 grid.

B.3 Additional Qualitative Analysis and Challenging Cases

The clinical descriptions generated by the VLM generally provide informative cues regarding anatomical locations and radiographic abnormalities. Nevertheless, local semantics may remain incomplete or uncertain for low-contrast lesions, regions with poorly defined boundaries, or cases involving overlapping radiographic findings. HiSTAR does not treat the generated descriptions as explicit diagnostic labels or supervisory ground truth; instead, they serve as auxiliary clinical cues that are jointly integrated with visual representations. Consequently, uncertainty in the semantic descriptions does not necessarily propagate into incorrect scanpath generation, as trajectory prediction remains grounded in both visual evidence and task-related semantic information. Representative challenging cases are presented in Fig. 10 and Fig. 11.

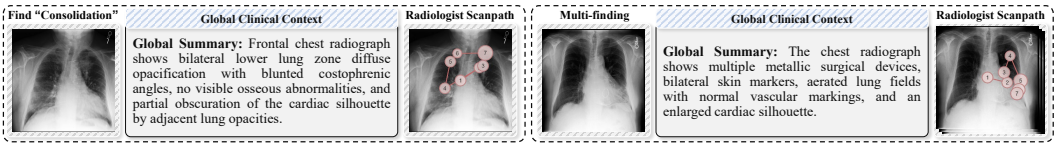


Figure 10: Representative challenging cases of global context on the GazeSearch dataset. The descriptions remain spatially and clinically relevant, although some subtle findings are summarized only coarsely.

In the first case, the global context does not explicitly assign the diagnostic term **[Consolidation]**, yet it preserves the major radiographic evidence, including diffuse bilateral lower-lung opacities, blunting of the costophrenic angles, and partial obscuration of the cardiac silhouette by adjacent pulmonary opacities. The second case contains multiple task labels, including **[Cardiomegaly]**, **[Pleural Other]**, and **[Support Devices]**. The generated global description accurately identifies the metallic devices and enlarged cardiac silhouette, while characterizing the subtle pleural abnormalities less comprehensively. Collectively, these examples illustrate that global context is intended to provide a holistic clinical representation rather than a complete diagnostic interpretation, allowing the corresponding task query to selectively emphasize information relevant to the target finding.

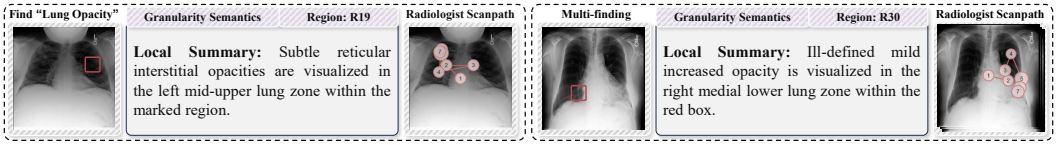


Figure 11: Representative challenging cases of granularity clinical semantics on the GazeSearch dataset.. Subtle or ambiguous regions retain task-relevant anatomical and spatial cues without being treated as diagnostic labels.

In R19 of the first case, the abnormality presents as a low-contrast reticular interstitial opacity. The local JSON assigns moderate confidence and characterizes the finding as "subtle", while preserving the correct anatomical localization in the left mid-to-upper lung. In the second case, R30 lies in the

medial right lower lung adjacent to the cardiomediastinal border. The VLM accurately localizes the region while assigning `uncertain` to `abnormality` and `ambiguous` to `boundary`, thereby avoiding an overly definitive interpretation when the visual evidence is insufficient. Collectively, these cases show that regional semantics can retain relatively stable anatomical and spatial cues even when the corresponding textual descriptions remain limited in precision.

C Detailed Architecture of HiSTAR

C.1 Hierarchical Semantic Adaptation

HiSTAR employs an ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) as its visual backbone. Each input image is resized to 224×224. The output of `layer1` provides a shallow, high-resolution feature map of size $256 \times 56 \times 56$, whereas `layer4` produces a deep, low-resolution feature map of size $2048 \times 7 \times 7$. Both feature maps are projected to a unified hidden dimension of $C = 384$ using 1×1 convolutions. Offline-generated 128-dimensional text embeddings are linearly projected into the model feature space.

For text embedding, the semantic relation encoder first augments each token with a fixed two-dimensional sinusoidal positional encoding and subsequently applies a single-layer Transformer Encoder with a hidden dimension of 384, 4 attention heads, a feed-forward dimension of 512, and a dropout rate of 0.1. For Eq. 13,

$$\widetilde{\mathbf{F}}_t^l \leftarrow \Phi_{t\leftarrow v}\left(\mathbf{F}_t^l, \mathbf{F}_v^h\right), \quad \widetilde{\mathbf{F}}_v^h \leftarrow \Phi_{v\leftarrow t}\left(\mathbf{F}_v^h, \mathbf{F}_t^l\right), \quad \mathbf{F}_r \leftarrow \Phi_f\left(\widetilde{\mathbf{F}}_v^h \| \widetilde{\mathbf{F}}_t^l\right), \tag{13}$$

both $\Phi_{t\leftarrow v}$ and $\Phi_{v\leftarrow t}$ are implemented using a single-layer Multi-Head Attention module with a hidden dimension of 384 and 4 attention heads. Residual connections and LayerNorm are applied in both directions. The resulting two 384-dimensional representations are concatenated into a 768-dimensional feature and then projected to 384 dimensions by the 1×1 convolution in Φ_f. The global mapping Φ_g is implemented as a linear layer followed by a residual addition with the original task query.

For spatial attention localization, the fused feature is first projected from 384 to 96 channels using a 1×1 convolution. Three transposed-convolution layers are then applied to progressively reduce the channel dimension to 48 while upsampling the feature to 56×56. After GroupNorm normalization, the resulting feature is concatenated with the 384-channel shallow visual feature, forming a 432-channel joint representation. The task query is projected through a linear transformation followed by ReLU and normalization, and is then matched with each spatial location using scaled dot-product similarity. The resulting Softmax probability $\mathbf{p}$ is further used to modulate the joint feature. Finally, the transposed-convolution mapping Φ_M progressively reduces the channel dimension and decodes the feature into a static diagnostic attention map of size 224×224.

C.2 Spatial-guided Temporal Decoding

Two-dimensional sine-cosine positional encodings with an embedding dimension of $C = 384$ are constructed for both the 56×56 shallow visual features and the 7×7 regional features. Historical fixation coordinates are first normalized and mapped onto the 56×56 feature grid, from which the corresponding shallow visual vectors are retrieved through discrete indexing. Each historical token is augmented with a learnable one-dimensional temporal embedding $E_{\text{time}}(i)$ and a type embedding $E_{\text{type}}(\text{hist})$, whereas the 49 regional tokens are assigned the type embedding $E_{\text{type}}(\text{region})$. Up to six historical fixations are retained, yielding a maximum memory sequence length of 55 tokens.

The self-attention mapping Φ_T is implemented as a three-layer Transformer Encoder with a hidden dimension of 384, 4 attention heads, a feed-forward dimension of 512, and a dropout rate of 0.1. After encoding, only the first 49 regional tokens are retained, enabling each regional representation to incorporate information from the accumulated fixation history. The static attention map $\mathbf{M}$ is adaptively average-pooled to a spatial resolution of 7×7. Meanwhile, the historical suppression map $\mathbf{M}_t^{\text{his}}$ is constructed by accumulating two-dimensional Gaussian responses centered at all previous fixation locations, followed by normalization with respect to the maximum response value. The initial values of λ_s and λ_h are set to 1.0 and -1.0, respectively, with Softplus-based parameterization used to enforce the corresponding sign constraints.

The fixation decoder, denoted by Φ_D, comprises a six-layer Transformer Decoder with a hidden dimension of 384, 4 attention heads, a feed-forward dimension of 512, and a dropout rate of 0.1. The

decoder query is initialized from the task representation refined by Φ_g, while the memory consists of the 49 regional tokens produced by Φ_T. The coordinate prediction head employs a three-layer MLP to transform the decoder output into a 384-dimensional query, which is subsequently matched against the 56×56 shallow visual features through dot-product similarity to obtain spatial logits. A 3×3 convolution is then applied, and the resulting response map is upsampled to 224×224 during scanpath generation. Both the duration and termination heads receive the 384-dimensional decoder representation together with the scalar history length and employ three-layer MLPs to produce scalar predictions. We apply softplus to the duration output to ensure non-negativity, while a sigmoid converts the termination logit into a probability.

D Ablation Studies of HiSTAR

D.1 Operator-Replacement Ablations on HSA and STD

Table 6: The ablation study on granularity semantic fusion on the GazeSearch dataset. MM. denotes the average MultiMatch score.

Setting	ScanMatch↑		MM.↑	SED↓	STDE↑
	w/o D.	w/ D.			
Local fusion→Add	0.3865	0.3558	0.8271	4.9346	0.8302
Local fusion→Concat	0.3881	0.3589	0.8286	4.9198	0.8320
HiSTAR (full)	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**

As shown in Table 6, we separately evaluate the contributions of local cross-modal fusion and temporal interaction. For the local fusion ablations, $\Phi_{t\leftarrow v}$, $\Phi_{v\leftarrow t}$, and Φ_f are jointly replaced with simpler aggregation operators. The *Add* variant performs element-wise addition between spatially corresponding visual and textual representations, whereas the *Concat* variant concatenates the two modalities and applies a linear projection to recover a 384-dimensional representation. Neither variant performs bidirectional conditional updating.

Table 7: Ablation on temporal interaction in STD on the GazeSearch dataset. MM. denotes the average MultiMatch score.

Setting	ScanMatch↑		MM.↑	SED↓	STDE↑
	w/o D.	w/ D.			
w/o Φ_T	0.3786	0.3462	0.8219	4.9027	0.8238
HiSTAR (full)	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**

As shown in Table 7, when Φ_T is removed, the fixation decoder directly receives the statically fused regional tokens, while historical fixation information is retained only through the suppression map. ScanMatch without duration decreases from 0.3942 to 0.3786, whereas SED increases from 4.8709 to 4.9027. The resulting degradation indicates that explicit temporal interaction between fixation history and current regional representations cannot be fully replaced by the static spatial prior.

Table 8: Ablation study of the architecture of the fixation decoder Φ_D on the GazeSearch dataset. MM. denotes the average MultiMatch score.

Setting	ScanMatch↑		MM.↑	SED↓	STDE↑
	w/o D.	w/ D.			
$\Phi_D \rightarrow$ Linear	0.3695	0.3384	0.8165	5.0714	0.8168
Φ_D: 3 layers	0.3867	0.3560	0.8277	4.9412	0.8309
Φ_D: 6 layers	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**
Φ_D: 9 layers	0.3928	0.3649	0.8315	4.8867	0.8359

With HSA, Φ_T, the spatial attention biases, and all prediction heads fixed, we vary the number of Transformer Decoder layers in the Fixation Decoder Φ_D. Using the original six-layer model as the baseline, we construct a shallower three-layer variant and a deeper nine-layer variant. Across all configurations, the hidden dimension of 384, 4 attention heads, a feed-forward dimension of 512, a dropout rate of 0.1, and the cross-attention mechanism between the task query and regional memory are kept unchanged. In addition, the *Linear* variant removes both self-attention and cross-attention from the decoder, providing a stronger structural ablation for evaluating the contribution of interaction between the task query and spatial memory.

Table 8 shows that reducing the decoder depth from six to three layers causes a moderate performance drop, whereas replacing the decoder with a linear mapping results in a substantially larger degradation. This comparison helps distinguish the effect of model capacity from the importance of interaction between the task query and spatial memory. The three-layer variant still preserves cross-attention, while the Linear variant cannot adaptively select regional information based on the evolving search state. Increasing the decoder depth to nine layers yields performance comparable to the six-layer baseline, with no further improvement.

Table 9: Ablation study of spatial disruption for granularity semantics on the GazeSearch dataset. MM. denotes the average MultiMatch score.

Disruption Ratio ρ	**ScanMatch↑**		**MM.↑**	**SED↓**	**STDE↑**
	w/o D.	**w/ D.**			
0%	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**
25%	0.3911	0.3610	0.8308	4.8981	0.8352
50%	0.3870	0.3573	0.8287	4.9234	0.8324
75%	0.3824	0.3523	0.8264	4.9468	0.8302
100%	0.3788	0.3487	0.8248	4.9705	0.8280

D.2 Disruption of Granularity Semantics

We investigate whether the performance gains of HiSTAR genuinely depend on the spatial correspondence between local clinical semantics and visual regions, rather than merely benefiting from the presence of additional textual tokens. For each sample, a proportion $\rho \in \{0, 0.25, 0.50, 0.75, 1.00\}$ of the 49 local semantic tokens is randomly selected, and the selected tokens are permuted among their corresponding positions. The permutation is defined as:

$$\widetilde{\mathbf{F}}_t^l[i] = \mathbf{F}_t^l[\pi(i)], \qquad i \in \Omega_\rho, \tag{14}$$

where Ω_ρ denotes the set of disrupted regions and π represents a random permutation restricted to Ω_ρ.

The results show a progressive performance degradation as the disruption ratio increases. A 25% permutation produces only a minor decline, suggesting that local semantic redundancy and neighborhood-level contextual propagation can partially compensate for limited spatial mismatches. As the disruption ratio increases to 50%-75%, ScanMatch, MultiMatch, and STDE consistently decrease, whereas SED continues to increase. The trend indicates that increasingly misaligned regional semantics interfere with both fixation ordering and spatial transitions during visual search. Under 100% disruption, the model still retains the original visual features, Global Context, and task query, preventing complete performance collapse. Nevertheless, the consistent deterioration confirms that preserving region-level semantic alignment is important for effective cross-modal interaction.

We further examine whether granularity-level semantics must be specific to the corresponding image through a cross-image semantic replacement experiment at test time. A proportion $\rho \in 0, 0.25, 0.50, 0.75, 1.00$ of target images is selected, and the complete 7×7 Granularity Clinical Semantics representation, consisting of 49 tokens, is replaced with that from another image. The original ordering of all 49 regional tokens is preserved, and no within-image spatial permutation is introduced. Meanwhile, the visual features, Global Context, task query, ground-truth scanpath, and model parameters of each target image remain unchanged. Such an experimental design isolates the effect of semantic–image inconsistency while avoiding additional confounding factors arising from spatial permutation.

Table 10: Cross-image replacement ablation of granularity clinical semantics on the GazeSearch dataset. ρ denotes the proportion of target samples whose complete local semantic representation is replaced by that from another image.

Replacement Ratio ρ	ScanMatch↑		MM.↑	SED↓	STDE↑
	w/o D.	w/ D.			
0%	**0.3942**	**0.3664**	**0.8328**	**4.8709**	**0.8375**
25%	0.3889	0.3608	0.8311	4.9037	0.8318
50%	0.3827	0.3573	0.8295	4.9406	0.8304
75%	0.3744	0.3491	0.8278	4.9794	0.8252
100%	0.3701	0.3436	0.8262	5.0135	0.8227

As shown in Table 10, performance progressively deteriorates as the proportion of cross-image semantic replacement increases: ScanMatch, MultiMatch, and STDE decrease, whereas SED increases. Unlike the within-image disruption experiment, which directly breaks the correspondence between regional semantics and spatial locations, cross-image replacement preserves the internal organization of the donor semantic grid but removes its semantic correspondence with the target image. The two experiments therefore provide complementary evidence: within-image permutation evaluates the importance of spatial alignment, while cross-image replacement examines image-specific semantic consistency. Together, the results suggest that the benefit of Granularity Clinical Semantics arises not merely from additional textual information, but from clinically meaningful semantics that remain both spatially aligned and image-specific.