A World Model of Radiologist Reading for Medical Image Representation Learning

Yiwei Li[1*] **Zihao Wu**[1*] **Huaqin Zhao**[1] **Yifan Zhou**[1]
Chao Cao[2] **Dajiang Zhu**[2] **Tianming Liu**[1] **Lin Zhao**[3†]
[1]University of Georgia [2]University of Texas at Arlington
[3]New Jersey Institute of Technology

Abstract

Radiologist eye-tracking data provide a rich record of how experts search, compare, and accumulate evidence during image reading; yet, existing methods exploit this signal only partially, either as a static spatial prior or as an auxiliary prediction target decoupled from diagnosis. We propose GazeWorld, a medical imaging world model that treats the image as the world and the radiologist's fixation sequence as a trajectory through it. GazeWorld autoregressively predicts the latent representation of the next fixated patch from the preceding fixation tokens, while a spatial-completion branch covers unvisited regions. Expert gaze supplies behavioral supervision for the prediction order, and latent targets are derived from images without diagnostic labels. At inference, frozen representations support downstream tasks without requiring recorded gaze. GazeWorld achieves the highest diagnostic AUROC among the compared methods across all nine supervised settings on CheXpert, RSNA Pneumonia, and SIIM-ACR Pneumothorax, as well as the highest zero-shot accuracy on all three benchmarks. On GazeSearch, a supervised scanpath decoder trained on frozen GazeWorld features improves ScanMatch by over 16% and reduces SED by over 22% relative to LogitGaze-Med, although the encoder's pretraining objective does not predict gaze coordinates. Temporal-order ablations support learning expert gaze-guided evidence-acquisition patterns, without establishing recovery of radiologists' explicit diagnostic reasoning.

1 Introduction

Two challenges limit the clinical translation of medical imaging AI. Large-scale medical imaging datasets remain scarce because acquisition is bound by clinical access, cross-institutional sharing is curtailed by privacy regulations, and expert annotations are prohibitively expensive to obtain at scale [Irvin et al., 2019, Johnson et al., 2019, Rajpurkar et al., 2017]. Without sufficiently large and diverse training sets, models readily overfit to dataset-specific shortcuts and degrade under distribution shifts [Azizi et al., 2021, Sowrirajan et al., 2021]. Compounding this data bottleneck, most current models emit a diagnostic label without exposing the evidence behind the prediction [Chen et al., 2024, Zhou et al., 2023], leaving clinicians unable to assess whether the output reflects sound reasoning or a spurious shortcut and ultimately eroding trust required for clinical deployment.

Radiologist gaze offers a promising inroad into both challenges. Eye-tracking data encode expert prior knowledge about where diagnostically relevant evidence lies, providing a form of supervision that can reduce reliance on large labeled datasets; several studies have incorporated gaze as a spatial attention prior and demonstrated improved performance [Shentu and Al Moubayed, 2024, Wang et al., 2022a, Ma et al., 2024, Wang et al., 2025], though collapsing eye movements into static

*Equal contribution.
†Corresponding author: lin.zhao.1@njit.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

heatmaps discards the order in which evidence was gathered. Beyond this, gaze trajectories record the temporal sequence in which a radiologist searches, compares, and accumulates evidence, providing an observable behavioral trace of evidence acquisition, rather than a direct annotation of diagnostic reasoning. Recent work has shown that such trajectories can be predicted from chest radiographs alone [Pham et al., 2025, Lvov and Pershin, 2025], but the predicted scanpaths remain decoupled from the diagnosis. Both lines leave the temporal structure of radiologist evidence gathering unused for diagnosis.

We propose to close this gap by formulating a world model of radiologist evidence acquisition. When reading a medical image, a radiologist navigates a perceptual world defined by the image. At each fixation, the radiologist extracts local evidence and selects the next region of interest conditioned on prior observations, iterating until sufficient evidence supports a diagnosis. We formalize this as a world model that predicts the latent representation of each successive fixation, conditioned on the current fixation and accumulated evidence. Each fixation along the trajectory thus becomes a prediction target, yielding image-derived latent targets without diagnostic labels. We call this expert-gaze-guided predictive pretraining: gaze is external behavioral supervision, while the latent targets are self-generated. This objective may reduce downstream label requirements, but it neither eliminates the cost of expert gaze collection nor directly supervises diagnostic reasoning.

We instantiate this formulation as GazeWorld[3], a medical imaging world model that captures the sequential structure of expert reading through gaze-ordered representation prediction. A visual encoder maps the image into patch-level representations and an autoregressive predictor, following the joint-embedding predictive architecture (JEPA) [LeCun et al., 2022, Assran et al., 2023], processes the gaze-ordered fixation sequence to predict the latent representation of the next fixated patch from all previously visited ones. In parallel, a spatial-completion branch uses the full gaze context to predict the representations of all unvisited patches. At inference, diagnostic features are extracted from the image using a deterministic surrogate order, and a separately trained scanpath decoder can probe the frozen visual features; neither requires recorded gaze at test time.

We evaluate GazeWorld on chest radiograph diagnosis and scanpath prediction. For diagnosis, linear probes on frozen GazeWorld features achieve the highest AUROC among the compared methods across all nine supervised settings on CheXpert [Irvin et al., 2019], RSNA Pneumonia [Colak et al., 2021], and SIIM-ACR Pneumothorax [Zawacki et al., 2019], as well as the highest zero-shot accuracy on all three benchmarks. For scanpath prediction, a supervised decoder on frozen GazeWorld features improves ScanMatch by over 16% and reduces SED by over 22% relative to LogitGaze-Med [Lvov and Pershin, 2025]. The encoder is not pretrained to predict fixation coordinates; the downstream scanpath decoder does receive gaze supervision.

Our contributions are as follows:

- We introduce the concept of a gaze world model for medical imaging, which formulates the radiologist's reading process as a trajectory through a perceptual world defined by the image. This formulation uses expert temporal order as behavioral supervision for image-derived latent prediction.
- We instantiate this formulation as GazeWorld, which combines gaze-ordered autoregressive prediction in latent space with a spatial-completion branch for unvisited regions, shaping the representation to capture both diagnostic content and sequential reading structure.
- GazeWorld achieves state-of-the-art diagnostic accuracy across all nine supervised settings and the highest zero-shot accuracy on three chest radiograph benchmarks. A generic decoder on the same frozen features outperforms purpose-built scanpath predictors, showing that a single representation supports both diagnostic transfer and scanpath decoding. A controlled diagnostic order ablation is included in the main text; additional medical self-supervised baselines, backbone comparisons, and resolution results are reported in Appendix F.

[3]Code: `https://github.com/levyisthebest/gazeworld`.

2 Related Work

2.1 Gaze Supervision and Scanpath Prediction

Radiologist gaze has been used in medical AI mainly in two ways. The first uses gaze as spatial supervision for diagnosis or representation learning, typically by converting fixations into heatmaps, attention maps, or relation matrices [Wang et al., 2022a, Saab et al., 2021, Stember et al., 2019, Hsieh et al., 2024]. These signals provide weak but clinically meaningful localization cues, helping models attend to regions that radiologists consider diagnostically relevant. For example, EGMA [Ma et al., 2024] uses fixation–derived attention to guide fine–grained image–text alignment, while eCLIP [Kumar and Marttinen, 2024] and other gaze–guided methods use expert attention to regularize contrastive learning, region–aware prediction, or diagnostic reasoning. These approaches demonstrate that gaze can improve clinical relevance and reduce reliance on dense manual annotations. However, they mostly treat gaze as static spatial supervision, preserving where radiologists looked while discarding the temporal order of evidence gathering, comparison, and integration.

The second line models gaze as a scanpath prediction problem. GazeFormer [Mondal et al., 2023], HAT [Yang et al., 2024], MedGaze [Awasthi et al., 2025], GazeSearch [Pham et al., 2025], and LogitGaze–Med [Lvov and Pershin, 2025] predict fixation trajectories or dwell times from medical images. These methods explicitly preserve gaze order and evaluate human–like visual search, but scanpath prediction is usually treated as a standalone output task. In contrast, GazeWorld uses radiologist scanpaths to define the autoregressive prediction order for representation learning, so the temporal structure of expert search directly shapes the visual encoder.

2.2 World Models and Representation–Space Prediction

World models aim to learn predictive structure in latent space as observations accumulate [LeCun et al., 2022]. Instead of reconstructing pixels, JEPA–style methods predict abstract representations of unseen regions from visible context, which makes the objective less sensitive to low–level appearance noise and more focused on semantic structure [Assran et al., 2023, Bardes et al., 2023]. Autoregressive vision models provide another predictive formulation by modeling visual tokens sequentially, but their prediction order is usually fixed by design, such as raster scan or predefined token layouts [Chen et al., 2020, El-Nouby et al., 2024]. These predictive objectives are especially appealing for medical imaging, where diagnostic signals are subtle, spatially distributed, and often better captured in representation space than at the pixel level. Recent medical models such as CheXWorld [Yue et al., 2025] and US–JEPA [Radhachandran et al., 2026] show that representation–space prediction can improve medical visual features by modeling anatomy, global layout, and domain variation.

However, existing predictive structures are typically defined by random masks, block masks, crops, or raster order. These choices are convenient but not tied to clinical reading behavior. GazeWorld instead uses radiologist fixation sequences as the prediction structure, where previously fixated patches form the context and the next expert–fixated patch is the autoregressive target. This grounds representation–space prediction in the temporal process of diagnostic evidence accumulation.

A complementary approach, Self-Evolving Code-with-Image Reasoning [Yang et al., 2026], executes visual algorithms in code and uses reflection on failed programs to develop reusable skills. GazeWorld instead learns visual representations through gaze-ordered latent prediction, without executing code as a reasoning interface.

2.3 Medical Vision Representation Learning

Medical vision models rely on supervised pretraining, self–supervised learning, contrastive learning, and vision–language pretraining to reduce annotation cost and improve transfer [Irvin et al., 2019, Johnson et al., 2019, Rajpurkar et al., 2017, Sowrirajan et al., 2021, Azizi et al., 2021, Xiao et al., 2023]. Representative medical vision–language models include GLoRIA [Huang et al., 2021], MedCLIP [Wang et al., 2022b], BioViL [Bannur et al., 2023], MGCA [Wang et al., 2022c], MedKLIP [Wu et al., 2023], and PRIOR [Cheng et al., 2023], which learn transferable representations from image labels, unlabeled images, or image–text pairs.

Gaze-free medical methods also include MedMAE [Gupta et al., 2025], which uses masked reconstruction, and LVM-Med [Nguyen et al., 2023], which uses self-supervised graph matching.

Appendix F.2 compares these methods under a common linear-probing setting. GazeWorld is complementary to these pretraining paradigms. It can be applied on top of existing medical visual backbones, such as MGCA, EGMA, or MedCLIP, to further refine their representations with expert gaze dynamics. Thus, our contribution is not a new foundation pretraining pipeline from scratch, but a gaze-ordered world-model refinement objective that introduces expert gaze structure into visual encoders. Its absolute performance and gain depend on the initialization; additional non-medical and random initializations are evaluated in Appendix F.3.

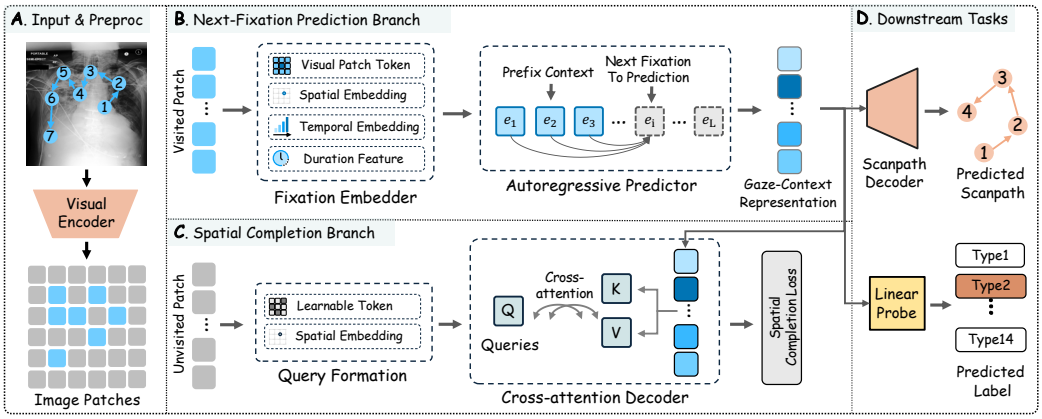


Figure 1: **GazeWorld overview.** After the chest radiograph and radiologist fixation sequence are processed in Part A, the image is converted into patch-level semantic tokens and the patch grid is divided into visited and unvisited regions. In Part B, visited patch tokens are combined with spatial, temporal, and duration information by the fixation embedder, and an autoregressive predictor performs latent-space prediction of the next expert-fixated representation from the prefix context. The resulting gaze-context representation is then used in Part C, where unvisited patch queries are formed from learnable tokens and spatial embeddings and completed through a cross-attention decoder. Finally, Part D evaluates the representations via scanpath decoding and diagnostic linear probing.

3 Method

GazeWorld casts the medical image as a perceptual world and the radiologist's fixation sequence as a trajectory through it. An autoregressive predictor processes the gaze-ordered fixation sequence and predicts the latent representation of the next fixated patch from all previously visited ones, using expert gaze to structure image-derived prediction targets. To exploit unvisited regions, a spatial-completion branch uses the full gaze context to predict all unvisited patch representations in parallel. The final pretraining objective combines next-fixation prediction with gaze-conditioned spatial completion.

3.1 Patch Representations and Fixation Sequence

Following the input and preprocessing stage in Fig. 1A, let $\mathbf{X}$ denote a medical image. An online visual encoder ϕ maps $\mathbf{X}$ into a set of patch-level semantic representations:

$$\mathbf{Z} = \phi(\mathbf{X}) = \{\mathbf{z}_p\}_{p=1}^{N}, \qquad \mathbf{z}_p \in \mathbb{R}^d, \tag{1}$$

where $\mathbf{Z}$ denotes the set of all patch-level representations, $\mathbf{z}_p$ is the d-dimensional semantic representation of the p-th image patch, and N is the number of image patches. The visual encoder is initialized from a pretrained model and adapted by the gaze-ordered objective.

Each image is paired with a fixation sequence $\mathcal{F} = \{f_1, \ldots, f_T\}$, where $f_t = (x_t, y_t)$ denotes the coordinates of the t-th fixation. We assign each fixation to its nearest image patch and merge only duplicate observations of the same patch within a time step. Revisits at different time steps remain separate tokens, distinguished by temporal rank and dwell duration. This yields a chronological sequence of visited patch indices, which may repeat, denoted by $\mathcal{S}$:

$$\mathcal{S} = (s_1, \ldots, s_L), \qquad s_i \in \{1, \ldots, N\}. \tag{2}$$

3.2 Next-Fixation Prediction

As illustrated in Fig. 1B, the next-fixation branch uses an autoregressive predictor to forecast the latent representation of each successive expert-fixated region from all preceding fixations. For each visited patch s_i in $\mathcal{S}$, we construct a fixation token by combining its visual representation with positional and temporal cues:

$$\mathbf{h}_i = \psi\left(\mathbf{z}_{s_i}, \mathbf{e}_{s_i}, i, \Delta_i\right), \tag{3}$$

where $\mathbf{e}_{s_i}$ is the spatial embedding of patch s_i, i the temporal rank of the fixation, Δ_i its dwell duration, and ψ the fixation embedder. Following the JEPA [LeCun et al., 2022, Assran et al., 2023], which learns by predicting target representations rather than reconstructing raw inputs, an autoregressive predictor g_θ with a causal attention mask receives the prefix of visited fixation tokens and predicts the representation of the next fixated patch:

$$\hat{\mathbf{z}}_{s_i} = g_\theta\left(\mathbf{h}_1, \ldots, \mathbf{h}_{i-1}\right), \qquad i = 2, \ldots, L. \tag{4}$$

Prediction targets come from a momentum target encoder $\bar{\phi}$, an exponential-moving-average (EMA) copy of the online visual encoder ϕ [Grill et al., 2020]. While ϕ is directly optimized by gradient descent, $\bar{\phi}$ is updated via stop-gradient after each optimization step, providing a stable representation space for the prediction objective:

$$\bar{\mathbf{z}}_{s_i} = \bar{\phi}(\mathbf{X})_{s_i}, \tag{5}$$

$$\bar{\theta}_\phi \leftarrow \tau\bar{\theta}_\phi + (1-\tau)\theta_\phi, \tag{6}$$

where θ_ϕ and $\bar{\theta}_\phi$ denote the parameters of the online visual encoder and the target encoder, respectively. The autoregressive loss is

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{L-1}\sum_{i=2}^{L} \mathrm{SmoothL1}\left(\mathrm{LN}(\hat{\mathbf{z}}_{s_i}), \mathrm{sg}(\bar{\mathbf{z}}_{s_i})\right), \tag{7}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization and $\mathrm{sg}(\cdot)$ stop-gradient. Following the JEPA principle, prediction operates in representation space rather than pixel space, encouraging contextual dependencies among expert-attended regions. The target is computed from the actual image; the predictor is not given the future fixation token. The objective estimates predictable latent structure, not the presence of an unseen abnormality. See Appendix F.1 for further discussion.

3.3 Spatial Completion

As shown in Fig. 1C, the spatial-completion branch complements the autoregressive objective, which supervises only the patches that appear in the radiologist's fixation sequence. However, unvisited regions can still be clinically informative. For example, a region may be skipped because it is normal, because it is peripheral to the suspected finding, or because the accumulated context is already sufficient. We therefore introduce a spatial-completion branch that assigns a representation-space target to every unvisited patch.

Let

$$\mathcal{U} = \{1, \ldots, N\} \setminus \{s_1, \ldots, s_L\} \tag{8}$$

denote the set of unvisited patches. A lightweight cross-attention decoder d_ξ predicts one representation for each unvisited patch. Queries combine a learnable mask token with the spatial embedding of the target patch; keys and values come from the gaze-context representation:

$$\hat{\mathbf{r}}_p = d_\xi\left(\mathbf{m} + \mathbf{e}_p, \mathbf{H}\right), \qquad p \in \mathcal{U}, \tag{9}$$

where $\mathbf{m}$ denotes a single learnable mask token shared across all unvisited patches, and $\mathbf{e}_p$ supplies its spatial position. The keys and values are

$$\mathbf{H} = g_\theta\left(\mathbf{h}_1, \ldots, \mathbf{h}_L\right) \tag{10}$$

the sequence of contextualized hidden states produced after processing the recorded fixation sequence, with $\mathbf{H} \in \mathbb{R}^{L\times 768}$ in the main configuration. It encodes fixation history and visual context, rather than a concatenation of predicted coordinates. Decoder details are given in Appendix B. The spatial-completion loss matches each prediction to the corresponding momentum target:

$$\mathcal{L}_{\mathrm{SC}} = \frac{1}{|\mathcal{U}|}\sum_{p\in\mathcal{U}} \mathrm{SmoothL1}\left(\mathrm{LN}(\hat{\mathbf{r}}_p), \mathrm{sg}\left(\mathrm{LN}(\bar{\mathbf{z}}_p)\right)\right). \tag{11}$$

This objective resembles masked representation prediction [Assran et al., 2023], but derives its context from the radiologist's reading trajectory rather than random visible patches. The model thus learns gaze-conditioned contextual representations for unvisited regions, without guaranteeing reconstruction of their unique visual content.

3.4 Pretraining Objective

The complete world-model pretraining objective combines autoregressive next-fixation prediction and spatial completion:

$$\mathcal{L}_{\mathrm{pre}} = \mathcal{L}_{\mathrm{AR}} + \lambda \mathcal{L}_{\mathrm{SC}}. \tag{12}$$

During pretraining, the online visual encoder is optimized jointly with the fixation embedder, autoregressive predictor, prediction head, and spatial-completion decoder. The target encoder receives no gradients and is updated only through EMA (6).

4 Experiments

4.1 Experimental Setup

We evaluate GazeWorld on diagnostic classification and scanpath prediction.

Diagnostic classification. We evaluate frozen backbone features of GazeWorld on CheXpert [Irvin et al., 2019], RSNA Pneumonia [Colak et al., 2021], and SIIM-ACR Pneumothorax [Zawacki et al., 2019] under 1%, 10%, and 100% label fractions, as well as zero-shot transfer on all three benchmarks. We compare against BioViL [Bannur et al., 2023], MedKLIP [Wu et al., 2023], MGCA [Wang et al., 2022c], GLoRIA [Huang et al., 2021], PRIOR [Cheng et al., 2023], MedCLIP [Wang et al., 2022b], CheXWorld [Yue et al., 2025], and EGMA [Ma et al., 2024].

Scanpath prediction. We pretrain on MIMIC-EYE [Hsieh et al., 2023] and probe on GazeSearch [Pham et al., 2025], a refined scanpath benchmark derived from MIMIC-EYE with more precise gaze-path annotations, under the ChestSearch protocol. For each backbone variant, we freeze the pretrained visual encoder and train only the same lightweight scanpath decoder on top of its patch-level features; performance differences therefore reflect scanpath-relevant information already encoded in the frozen representations. We compare against representative scanpath prediction methods, including general visual-search baselines GazeFormer [Mondal et al., 2023] and HAT [Yang et al., 2024], radiology-specific GazeSearch [Pham et al., 2025], and the recent LogitGaze / LogitGaze-Med models [Lvov and Pershin, 2025], with LogitGaze-Med serving as the strongest scanpath baseline.

All dataset statistics, baseline protocols, and split details are provided in Appendices A.1 and A.2.

4.2 Implementation Details

We use MedCLIP-Swin [Liu et al., 2021] as the visual encoder, processing chest radiographs at 224×224 resolution. The autoregressive predictor is an 8-layer causal Transformer with hidden dimension 768. GazeWorld is pretrained on MIMIC-EYE [Hsieh et al., 2023] for 15 epochs using AdamW [Loshchilov and Hutter, 2017] with learning rate 3×10^{-4}. The spatial-completion weight is set to $\lambda = 1.0$. At evaluation, the visual encoder and predictor are frozen; downstream tasks use only the resulting representations with no access to gaze data. Full architecture specifications, hyperparameters, and evaluation protocols are provided in Appendix B.

4.3 Downstream Classification

If the world model has captured the invariants of expert reading, its representations should also transfer to standard diagnostic tasks. We evaluate frozen backbone features for supervised and zero-shot classification on three external datasets that are not used during world-model pretraining.

Classification results. Tables 1 and 2 summarize supervised and zero-shot diagnostic transfer. In supervised classification, GazeWorld achieves the highest reported AUROC in all nine dataset–label settings. These are point-estimate comparisons, without claims of statistical significance. Under the

Table 1: Comparison of supervised classification AUROC on CheXpert [Irvin et al., 2019], RSNA [Colak et al., 2021], and SIIM-ACR [Zawacki et al., 2019] with 1%, 10%, and 100% of training labels. Results are point estimates. **Best** and second-best results are highlighted.

Method	CheXpert			RSNA			SIIM-ACR		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
BioViL [Bannur et al., 2023]	54.19	60.37	66.55	70.74	72.84	78.48	62.98	68.39	78.20
MedKLIP [Wu et al., 2023]	-	-	-	79.43	78.62	82.76	74.50	78.27	84.67
MGCA [Wang et al., 2022c]	67.12	76.10	81.57	80.41	85.67	88.24	86.14	88.13	91.84
GLoRIA [Huang et al., 2021]	67.40	77.37	81.74	81.90	81.24	88.57	-	-	-
PRIOR [Cheng et al., 2023]	65.51	70.70	79.11	79.74	79.72	83.95	81.96	82.76	85.33
CheXWorld [Yue et al., 2025]	62.81	80.03	85.39	78.30	84.12	88.66	62.31	81.54	91.34
MedCLIP [Wang et al., 2022b]	74.65	82.88	85.58	82.31	84.16	87.86	86.09	87.41	91.46
EGMA [Ma et al., 2024]	71.21	79.79	85.45	82.49	84.73	88.09	87.17	87.91	88.98
Ours	**78.37**	**83.41**	**87.61**	**83.27**	**85.75**	**90.15**	**87.85**	**90.48**	**94.27**

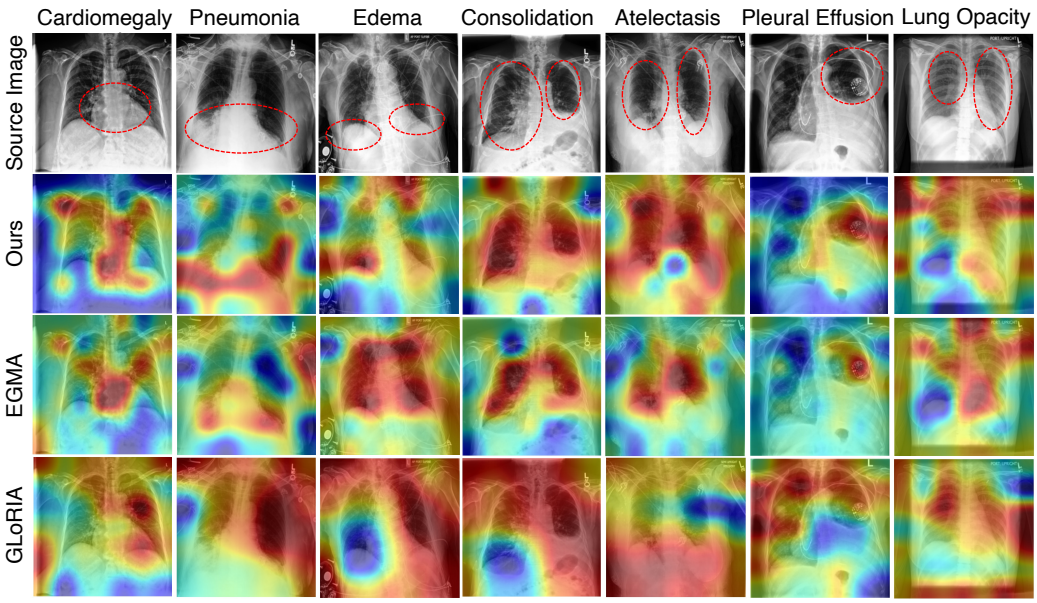


Figure 2: Grad-CAM [Selvaraju et al., 2017] attention visualizations across seven pathologies, comparing the source radiograph, GazeWorld, EGMA [Ma et al., 2024], and GLoRIA [Huang et al., 2021]. Red dashed circles on the source radiographs highlight pathology-relevant regions used for visual reference. All maps are generated with the same Grad-CAM protocol (Appendix C).

controlled comparison with EGMA, where the Swin backbone, MedCLIP initialization, gaze data, and downstream protocol are matched, GazeWorld improves CheXpert 1% AUROC from 71.21 to 78.37. It also outperforms CheXWorld [Yue et al., 2025] in all settings, with the largest gains in low-label regimes. In zero-shot evaluation, GazeWorld obtains the highest accuracy on all three datasets and the highest F1 on CheXpert and SIIM-ACR. Figure 2 provides qualitative Grad-CAM visualizations under the EGMA protocol (Appendix C), showing that GazeWorld produces more concentrated activations over clinically relevant anatomy than EGMA and GLoRIA. These visualizations are qualitative and do not establish localization accuracy or faithful explanations. Training and evaluation details are summarized in Appendix B.

Additional medical SSL comparisons. In separate linear-probing runs, GazeWorld has the highest point estimate in all nine dataset–label settings against CheXWorld, MedMAE, and LVM-Med (Appendix F.2). At 100% CheXpert labels, AUROC is 87.6, compared with 85.5, 76.6, and 73.1

Table 2: Zero-shot classification on CheXpert 5×200 [Irvin et al., 2019], RSNA [Colak et al., 2021], and SIIM-ACR [Zawacki et al., 2019]. AUROC, Accuracy, and F1 are computed without task-specific fine-tuning. **Best** and second-best results are highlighted.

Method	CheXpert 5×200		RSNA		SIIM-ACR	
	AUROC↑	Acc.↑ / F1↑	AUROC↑	Acc.↑ / F1↑	AUROC↑	Acc.↑ / F1↑
CLIP [Radford et al., 2021]	15.50	19.60 / 10.99	21.67	23.46 / 19.87	42.13	46.11 / 40.13
GLoRIA [Huang et al., 2021]	49.18	51.10 / 47.25	60.77	61.53 / **60.09**	–	- / -
MGCA [Wang et al., 2022c]	42.53	43.39 / 39.67	59.19	60.82 / 59.57	36.60	38.45 / 34.95
MedCLIP [Wang et al., 2022b]	56.63	57.40 / 55.72	58.08	62.17 / 40.06	56.17	56.89 / 55.44
EGMA [Ma et al., 2024]	56.09	57.15 / 55.03	57.46	66.85 / 43.08	59.79	63.24 / 61.04
Ours	**58.13**	**59.42 / 56.85**	**62.84**	**70.18** / 47.31	**64.29**	**66.59 / 63.26**

for these baselines (LVM-Med ResNet-50). Their different initializations and pretraining data limit causal attribution to gaze supervision.

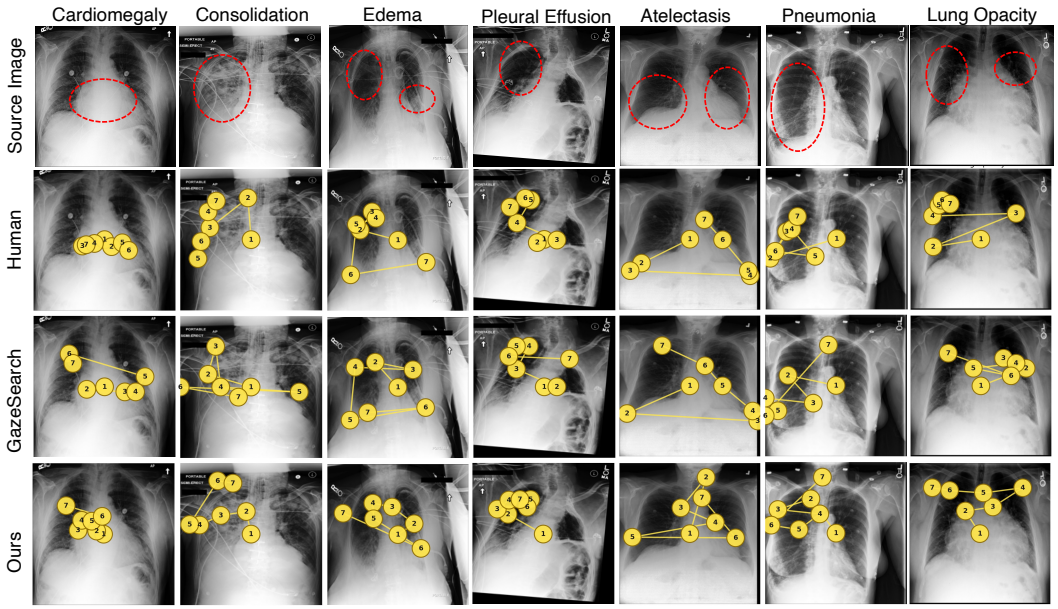


Figure 3: Qualitative scanpath comparison across seven pathologies. The first row shows the source radiographs, with red dashed circles marking pathology-relevant regions. We compare human scanpaths, GazeSearch predictions [Pham et al., 2025], and trajectories decoded from frozen GazeWorld representations using the same supervised decoder.

4.4 Scanpath Prediction

A central claim of this work is that world-model pretraining on fixation sequences produces representations that encode the sequential structure of expert reading. We test this claim with a controlled probing experiment. For each backbone variant, we freeze the pretrained visual encoder and train only a lightweight scanpath decoder on top of its patch-level features (Appendix D). The decoder architecture, optimization settings, and training data are kept identical across variants; the only variable is the frozen representation. Thus, performance differences reflect how much scanpath-relevant information is already encoded in the learned features.

Table 3 shows that all three variants of GazeWorld outperform the previous state-of-the-art LogitGaze-Med on the main scanpath metrics. The best variant improves ScanMatch [Cristino et al., 2010] by 16.7% (0.489 vs. 0.419), reduces SED [Foulsham and Underwood, 2008] by 22.4% (3.63 vs. 4.68), and raises STDE [Wang et al., 2011] by over 6% relative to LogitGaze-Med (CheX). Both LogitGaze-Med and our downstream decoder use gaze supervision. Our encoder is trained with

Table 3: GazeSearch scanpath probing results. ScanMatch, STDE, and MultiMatch scores are higher-better, while SED is lower-better. **Best** and second-best results are highlighted.

Method	Scanpath			MultiMatch		
	SM↑	SED↓	STDE↑	Vec.↑	Dir.↑	Pos.↑
GazeFormer [Mondal et al., 2023]	0.280	5.11	0.801	0.902	0.644	0.803
HAT [Yang et al., 2024]	0.296	5.07	0.800	0.909	0.649	0.825
GazeSearch [Pham et al., 2025]	0.351	4.89	0.809	0.917	0.679	0.829
LogitGaze [Lvov and Pershin, 2025]	0.328	5.07	0.810	0.882	0.643	0.809
LogitGaze-Med (Res) [Lvov and Pershin, 2025]	0.416	4.68	0.852	0.935	0.650	0.823
LogitGaze-Med (CheX) [Lvov and Pershin, 2025]	0.419	4.68	0.855	0.938	0.651	0.823
Ours (MedCLIP)	0.479	3.74	0.909	0.941	0.649	0.837
Ours (Res)	0.484	3.66	**0.912**	**0.966**	0.678	**0.917**
Ours (CheX)	**0.489**	**3.63**	0.911	**0.966**	**0.686**	0.915

latent prediction rather than coordinate regression; these gains suggest that gaze-ordered world-model pretraining encodes scanpath-relevant structure into the learned representation. The same table further reports MultiMatch [Jarodzka et al., 2010] results, with consistent gains across vector, direction, and position.

Figure 3 provides qualitative support for the probing results. Across seven pathologies, scanpaths decoded from GazeWorld features more closely follow radiologist reading trajectories and attend to pathology-relevant anatomy, whereas GazeSearch produces more diffuse fixation sequences.

4.5 Ablation Studies

Gaze ordering and objectives. Table 4 shows the benefit of chronological gaze order: replacing expert scanpaths with raster-scan or random order substantially degrades zero-shot AUROC across all datasets, indicating that the gain is not merely due to ordered patch exposure. MedCLIP initialization performs best, suggesting that gaze-ordered prediction benefits from a semantically aligned medical visual backbone. Finally, AR-only already gives strong performance, while adding spatial completion further improves results; SC-only is weaker, suggesting that predicting unvisited anatomy without the autoregressive fixation context provides less effective supervision, and that completion is most beneficial when conditioned on chronological expert fixation dynamics.

Backbone initialization. Table 5 further evaluates whether GazeWorld can improve different medical visual backbones. Across MGCA, EGMA, and MedCLIP initializations, adding GazeWorld consistently improves zero-shot performance on the external benchmarks, showing that the proposed gaze-ordered world-model pretraining is not tied to a single backbone. Additional 100%-label probes include DINOv3, CheXWorld, and a random Swin initialization (Appendix F.3). All tested initializations improve, but both gain size and absolute performance vary. Table 6 adds a controlled diagnostic order ablation, and Appendix F.2 compares gaze-free medical SSL baselines. These results support the tested refinement setting rather than universal effectiveness or recovery of explicit reasoning.

Controlled diagnostic order ablation. Table 6 tests chronological expert gaze against raster and random orders with the training setting and gaze-selected patch observations held fixed. Real gaze exceeds raster order by 5.20, 1.54, and 8.28 AUROC points on CheXpert 5×200, RSNA, and SIIM-ACR, respectively. The gains over random order are 7.55, 4.69, and 12.30 points. Thus, expert temporal order contributes beyond selecting relevant patches, although the experiment does not establish recovery of explicit clinical reasoning.

Table 4: Zero-shot ablations of fixation order and objective components. AR denotes next-fixation prediction; SC denotes spatial completion. All metrics are computed without task-specific fine-tuning.

Variant	CheXpert 5×200		RSNA		SIIM-ACR	
	AUROC↑	Acc.↑ / F1↑	AUROC↑	Acc.↑ / F1↑	AUROC↑	Acc.↑ / F1↑
Raster-scan Order	52.18	55.21 / 51.58	50.39	62.46 / 42.30	53.50	56.22 / 50.16
Random Order	52.25	56.08 / 49.13	51.58	61.53 / 42.68	55.06	57.02 / 54.34
AR Only	57.78	58.86 / 55.91	59.11	68.07 / 44.15	62.10	64.88 / 61.94
SC Only	54.67	54.41 / 53.06	51.25	60.63 / 40.03	52.59	53.72 / 51.37
Ours	58.13	59.42 / 56.85	62.84	70.18 / 47.31	64.29	66.59 / 63.26

Table 5: Effect of backbone initialization on zero-shot classification. AUROC, Accuracy, and F1 are reported on three external datasets.

Backbone	CheXpert 5×200			RSNA			SIIM-ACR		
	AUROC↑	Acc.↑	F1↑	AUROC↑	Acc.↑	F1↑	AUROC↑	Acc.↑	F1↑
MGCA	42.53	43.39	39.67	59.19	60.82	59.57	36.60	38.45	34.95
+GazeWorld	47.41	48.31	46.07	60.76	61.46	59.98	41.07	43.56	39.62
Δ	+4.88	+4.92	+6.40	+1.57	+0.64	+0.41	+4.47	+5.11	+4.67
EGMA	56.09	57.15	55.03	57.46	66.85	43.08	59.77	63.24	61.04
+GazeWorld	56.42	58.30	55.45	58.50	67.16	44.89	60.83	64.01	61.33
Δ	+0.33	+1.15	+0.42	+1.04	+0.31	+1.81	+1.06	+0.77	+0.29
MedCLIP	56.63	57.40	55.72	58.09	62.17	40.07	56.17	56.89	55.44
+GazeWorld	58.13	59.42	56.85	62.84	70.18	47.31	64.29	66.59	63.26
Δ	+1.50	+2.02	+1.13	+4.75	+8.01	+7.24	+8.12	+9.70	+7.82

Table 6: Controlled diagnostic order ablation: AUROC (%) with 100% labeled linear probing. These ablation runs are separate from Table 1.

Prediction order	CheXpert 5×200	RSNA	SIIM-ACR
Chronological expert gaze	**87.59**	**90.15**	**94.26**
Raster	82.39	88.61	85.98
Random	80.04	85.46	81.96

5 Conclusion

GazeWorld combines next-fixation prediction and spatial completion to transfer expert visual-search structure into image representations. Diagnostic and scanpath results support this gaze-guided refinement objective, without establishing recovery of explicit clinical reasoning.

Limitations. Evaluation is limited to 2D chest radiographs and one gaze corpus; GazeSearch shares that source. External diagnostic transfer does not establish robustness to new readers, institutions, or modalities. Patch discretization loses subpatch precision, and results depend on initialization. Qualitative heatmaps do not establish localization accuracy or faithful explanations; point estimates do not establish statistical significance. Multi-reader validation, quantitative localization, and CT/MRI extensions remain future work (Appendix F.5).

Broader impact. Gaze-guided representations may support label-efficient medical imaging, but reader bias, misleading heatmaps, and diagnostic errors could harm patients. Clinical use requires independent validation, subgroup assessment, clinician oversight, and protection of patient and eye-tracking data.

Acknowledgments and Disclosure of Funding

The authors thank Chong Ma of the School of Computing and Artificial Intelligence, Southwest Jiaotong University (machong@swjtu.edu.cn), for his assistance with data and algorithms.

The authors received no third-party funding or financial support for this work. The authors declare no competing interests.

References

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

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.

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

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Junjie Shentu and Noura Al Moubayed. Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5212–5221, 2024.

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, 2022a.

Chong Ma, Hanqi Jiang, Wenting Chen, Yiwei Li, Zihao Wu, Xiaowei Yu, Zhengliang Liu, Lei Guo, Dajiang Zhu, Tuo Zhang, et al. Eye-gaze guided multi-modal alignment for medical representation learning. *Advances in Neural Information Processing Systems*, 37:6126–6153, 2024.

Sheng Wang, Zihao Zhao, Zhenrong Shen, Bin Wang, Qian Wang, and Dinggang Shen. Improving self-supervised medical image pre-training by early alignment with human eye gaze information. *IEEE Transactions on Medical Imaging*, 44(10):4063–4072, 2025.

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. IEEE, 2025.

Dmitry Lvov and Ilya Pershin. From human attention to diagnosis: Semantic patch-level integration of vision-language models in medical imaging. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

Andrew Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. Kaggle, 2019. Available at: `https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation`.

Khaled Saab, Sarah M Hooper, Nimit S Sohoni, Jupinder Parmar, Brian Pogatchnik, Sen Wu, Jared A Dunnmon, Hongyang R Zhang, Daniel Rubin, and Christopher Ré. Observational supervision for medical image classification using gaze data. In *International conference on medical image computing and computer-assisted Intervention*, pages 603–614. Springer, 2021.

Joseph N Stember, Haydar Celik, E Krupinski, Peter D Chang, Simukayi Mutasa, Bradford J Wood, A Lignelli, Gul Moonis, Lawrence H Schwartz, Sachin Jambawalikar, et al. Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of digital imaging*, 32(4): 597–604, 2019.

Chihcheng Hsieh, André Luís, José Neves, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Joaquim Jorge, and Catarina Moreira. Eyexnet: Enhancing abnormality detection and diagnosis via eye-tracking and x-ray fusion. *Machine Learning and Knowledge Extraction*, 6(2):1055–1071, 2024.

Yogesh Kumar and Pekka Marttinen. Improving medical multi-modal contrastive learning with expert annotations. In *European Conference on Computer Vision*, pages 468–486. Springer, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1441–1450, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024.

Akash Awasthi, Anh Mai Vu, Ngan Le, Zhigang Deng, Supratik Maulik, Rishi Agrawal, Carol C Wu, and Hien Van Nguyen. Modeling radiologists' cognitive processes using a digital gaze twin to enhance radiology training. *Scientific reports*, 15(1):13685, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. Chexworld: Exploring image world modeling for radiograph representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20778–20788, 2025.

Ashwath Radhachandran, Vedrana Ivezić, Shreeram Athreya, Ronit Anilkumar, Corey W Arnold, and William Speier. Us-jepa: A joint embedding predictive architecture for medical ultrasound. *arXiv preprint arXiv:2602.19322*, 2026.

Tianze Yang, Liang Wu, Ruitong Sun, Yucheng Shi, Yanqiao Wang, Mayank Darbari, Ninghao Liu, Jin Sun, and Liangjie Hong. Self-evolving code-with-image reasoning. *arXiv preprint arXiv:2608.11292*, 2026.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3588–3600, 2023.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951, 2021.

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*, pages 3876–3887, 2022b.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15016–15027, 2023.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022c.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21361–21371, 2023.

Anubhav Gupta, Islam Osman, Mohamed S. Shehata, W. John Braun, and Rebecca E. Feldman. MedMAE: A self-supervised backbone for medical imaging tasks. *Computation*, 13(4):88, 2025. doi: 10.3390/computation13040088.

Duy M. H. Nguyen, Hoang Nguyen, Nghiem T. Diep, Tan N. Pham, Tri Cao, Binh T. Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, Daniel Sonntag, and Mathias Niepert. LVM-Med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. In *Advances in Neural Information Processing Systems*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Chihcheng Hsieh, Chun Ouyang, Jacinto C Nascimento, Joao Pereira, Joaquim Jorge, and Catarina Moreira. Mimic-eye: Integrating mimic datasets with reflacx and eye gaze for multimodal deep learning applications. *PhysioNet (version 1.0. 0)*, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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.

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.

Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *CVPR 2011*, pages 441–448. IEEE, 2011.

Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 211–218, 2010.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

A Baseline Protocol Details

A.1 Dataset Details

MIMIC-EYE. MIMIC-EYE [Hsieh et al., 2023] is an eye-tracking extension of the MIMIC-CXR database [Johnson et al., 2019]. It records radiologist fixation sequences from $3{,}032$ frontal chest radiographs during routine clinical reading sessions using a Tobii Pro Nano eye tracker (sampling rate 60 Hz). Each record contains a temporally ordered sequence of fixation coordinates (x_t, y_t) with associated dwell durations. We split the data into $2{,}729$ / 151 / 152 train/val/test following the official partition. Per-image fixation sequences average 47 fixation points; after merging duplicate within-step observations at patch level, temporal revisits remain in the sequence used for world-model pretraining.

GazeSearch (ChestSearch protocol). GazeSearch [Pham et al., 2025] provides task-directed eye-tracking data: radiologists search for specific pathologies (13 finding categories) on chest radiographs. The ChestSearch split contains $3{,}870$ training and 488 test samples, each consisting of a radiograph, a target pathology label, and a ground-truth fixation sequence (up to 7 fixations).

CheXpert. CheXpert [Irvin et al., 2019] is a large-scale chest radiograph dataset containing $224{,}316$ images from $65{,}240$ patients at Stanford Hospital, labeled for 14 observations via an automated NLP labeler applied to radiology reports. The main supervised probing protocol uses 14-class macro AUROC under 1%, 10%, and 100% label fractions, as detailed below. For zero-shot evaluation, we follow the CheXpert 5×200 protocol [Ma et al., 2024]: 200 images per class, 5 pathologies.

RSNA Pneumonia. The RSNA [Colak et al., 2021] Pneumonia Detection Challenge dataset provides $26{,}684$ frontal chest radiographs with bounding-box annotations for pneumonia opacities, curated from the NIH ChestX-ray14 collection. We use the binary pneumonia/normal classification task under 1%, 10%, and 100% label fractions.

SIIM-ACR Pneumothorax. The SIIM-ACR Pneumothorax Segmentation dataset [Zawacki et al., 2019] contains $12{,}047$ chest radiographs with pixel-level pneumothorax masks. We convert it to a binary classification task (pneumothorax-present vs. normal) and evaluate under 1%, 10%, and 100% label fractions.

A.2 Baseline Comparison Protocol

Table 7 summarizes the backbone architecture, pretraining data, and result provenance for every baseline in the main comparison tables.

Classification baselines. Classification baseline point estimates for BioViL, MedKLIP, MGCA, GLoRIA, PRIOR, MedCLIP, and EGMA in Tables 1 and 2 are taken from EGMA [Ma et al., 2024], which re-evaluated every method under a single downstream protocol. The EGMA protocol fixes the following variables across all methods: (i) train/val/test splits, (ii) downstream head, (iii) optimizer and schedule, (iv) number of fine-tuning epochs, and early-stopping criterion, and (v) metric computation. The protocol does not harmonize backbone architecture: BioViL, MedKLIP, MGCA, GLoRIA, and PRIOR use ResNet-50, while MedCLIP, EGMA, and our method use Swin-Tiny. Therefore, comparisons against ResNet-50-based baselines include both objective and backbone differences.

Controlled comparison (Ours vs. EGMA). The cleanest comparison in our evaluation is Ours vs. EGMA. Both methods share the same Swin backbone, the same MedCLIP initialization weights, the same MIMIC-EYE fixation data, and the same downstream evaluation protocol. The only variable is the pretraining objective applied on top of MedCLIP: EGMA uses gaze-guided contrastive alignment, while we use the GazeWorld world-model objective. Performance differences between these two rows in Tables 1–2 therefore isolate the contribution of the pretraining objective with minimal confounds.

Scanpath baselines. Scanpath prediction baselines in Table 3 are evaluated under the ChestSearch protocol. We use the original paper results for LogitGaze-Med use the reported HAT result from GazeSearch, and evaluate GazeFormer, GazeSearch, and LogitGaze using their official implementations under the same protocol. Our scanpath decoder uses the same evaluation protocol (7 fixations,

the same test split of 488 samples, and the same metrics). For the scanpath probe, the decoder architecture and training procedure are held fixed across backbone variants.

Resolution and backbone notes. For classification (Tables 1 and 2), all methods use 224×224 input resolution. For scanpath prediction (Table 3), our decoder and the primary baselines operate at 448×448; GazeFormer and HAT use 224×224 but are included for completeness as earlier methods.

B Implementation Details

Backbone. We use the MedCLIP Swin image encoder [Liu et al., 2021] as the online visual backbone. The encoder has 4 stages, embedding dimension $C = 96$, window size 7, patch size 4, and output dimension $d = 768$. Unless otherwise specified, chest radiographs are resized to 224×224 for representation learning, yielding patch-level semantic tokens

$$\mathbf{Z} = \phi(\mathbf{X}) = \{\mathbf{z}_p\}_{p=1}^{N}, \qquad \mathbf{z}_p \in \mathbb{R}^{768}. \tag{13}$$

During world-model pretraining, the online visual encoder is updated jointly with the gaze-conditioned modules. A separate momentum target encoder provides stop-gradient representation-space targets.

Fixation preprocessing. Each eye-tracking record provides a temporally ordered sequence of fixation coordinates with associated dwell durations. Fixation coordinates are normalized to the image frame and mapped to the nearest patch index. Only duplicate observations within the same time step are merged. Visits to the same patch at different time steps retain separate temporal-rank and dwell-duration features, yielding a chronological sequence of potentially repeated patch indices $\mathcal{S} = (s_1, \ldots, s_L)$. This sequence defines the causal prediction order used by the autoregressive world model.

Fixation embedder. For each visited patch s_i, the fixation embedder combines the visual patch token $\mathbf{z}_{s_i}$ with three fixation-specific signals: a two-dimensional spatial embedding of the patch location, a learned temporal-rank embedding for the fixation order i, and a projected dwell-duration feature. The resulting fixation token is used as input to the causal predictor.

Autoregressive predictor. The autoregressive predictor is an 8-layer causal Transformer with 12 attention heads, head dimension 64, hidden dimension $d = 768$, and a lower-triangular causal attention mask. It receives previously fixated tokens and predicts the semantic representation of the next fixated patch. The prediction head is a two-layer MLP with GELU activation.

Momentum target encoder. To provide stable representation-space targets, we maintain a target encoder $\bar{\phi}$ as an exponential-moving-average copy of the online visual encoder ϕ. The target encoder is evaluated under no-gradient computation and is not directly updated by back-propagation. After each optimizer step, its parameters are updated as

$$\bar{\theta}_\phi \leftarrow \tau \bar{\theta}_\phi + (1 - \tau)\theta_\phi, \tag{14}$$

where θ_ϕ and $\bar{\theta}_\phi$ denote the parameters of the online and target visual encoders, respectively. The EMA momentum follows a cosine schedule from 0.998 to 1.0.

Spatial-completion decoder. The spatial-completion branch is a lightweight 2-layer cross-attention decoder with hidden dimension $d = 768$. Let $\mathcal{U} = \{1, \ldots, N\} \setminus \{s_1, \ldots, s_L\}$ denote the set of unvisited patches. For each $p \in \mathcal{U}$, the decoder forms a query by adding a single shared learnable mask token to the position-specific learned spatial embedding of patch p. The keys and values are the gaze-context tokens produced by the causal predictor. The decoder outputs one predicted representation $\hat{\mathbf{r}}_p$ for each unvisited patch, which is matched to the corresponding momentum-target representation $\bar{\mathbf{z}}_p$. Each of the two cross-attention blocks uses 12 heads of dimension 64, RMSNorm, and a SwiGLU feed-forward network with hidden dimension 2048. Query, key, value, and output attention projections are bias-free. A linear output head maps 768-dimensional hidden states to the per-patch target feature dimension.

Pretraining objective. The world-model pretraining objective combines autoregressive next-fixation prediction and spatial completion:

$$\mathcal{L} = \mathcal{L}_{\mathrm{AR}} + \lambda \mathcal{L}_{\mathrm{SC}}. \tag{15}$$

In all main experiments, we set the spatial-completion weight to $\lambda = 1.0$. The autoregressive loss is

$$\mathcal{L}_{\mathrm{AR}} = \frac{1}{L-1} \sum_{i=2}^{L} \mathrm{SmoothL1}\left(\mathrm{LN}(\hat{\mathbf{z}}_{s_i}), \mathrm{sg}(\bar{\mathbf{z}}_{s_i})\right), \tag{16}$$

and the spatial-completion loss is

$$\mathcal{L}_{\mathrm{SC}} = \frac{1}{|\mathcal{U}|} \sum_{p \in \mathcal{U}} \mathrm{SmoothL1}\left(\mathrm{LN}(\hat{\mathbf{r}}_p), \mathrm{sg}\left(\mathrm{LN}(\bar{\mathbf{z}}_p)\right)\right). \tag{17}$$

Here, $\mathrm{LN}(\cdot)$ denotes layer normalization and $\mathrm{sg}(\cdot)$ denotes stop-gradient.

Optimization. We train the world model on MIMIC-EYE using AdamW [Loshchilov and Hutter, 2017] with learning rate 3×10^{-4}, weight decay 0.04, batch size 32, automatic mixed precision, and 15 training epochs. The online MedCLIP-Swin encoder, fixation embedder, causal predictor, autoregressive prediction head, and spatial-completion decoder are optimized jointly. The target encoder receives no gradients and is updated only through EMA.

Scanpath decoder. GazeSearch train split: $3{,}870$ samples. AdamW, learning rate 5×10^{-4}, weight decay 0.01, cosine annealing ($\eta_{\text{min}} = 10^{-6}$, $T_{\text{max}} = 30$), batch size 16, 30 epochs. Best checkpoint selected by validation mean Euclidean distance; evaluation on 488 test samples.

Linear probing. We evaluate frozen pretrained representations on CheXpert (14-class), RSNA Pneumonia, and SIIM-ACR Pneumothorax at 1%, 10%, and 100% label fractions. For each image we extract a 1536-d feature by concatenating two halves: (i) the mean-pooled patch tokens of the online MedCLIP-Swin encoder ϕ, and (ii) the d-dimensional output of a learnable readout token prepended to a deterministic raster-scan surrogate fixation sequence over the N patches and processed by the frozen causal predictor, followed by a learned linear projection back to d. No gaze annotations are used at probing time. Features are standardised per dimension (zero mean, unit variance) and fed to an ℓ_2-regularised logistic-regression classifier (LBFGS, C=1, 2000 iterations); training fractions are sampled with a fixed seed. We report 14-class macro AUROC on CheXpert and binary AUROC on RSNA / SIIM-ACR. The backbone and predictor receive no gradient at probing time.

C Attention Heatmap Protocol

The attention heatmaps in Figure 2 are generated using Grad-CAM [Selvaraju et al., 2017], following the same protocol as EGMA [Ma et al., 2024] to ensure visual comparability. For each image–pathology pair: (i) the frozen backbone used for downstream evaluation processes the image at 224×224; (ii) we compute the gradient of the target pathology logit with respect to the feature maps of the final convolutional stage (Swin-Tiny stage 4 for our method and EGMA; the corresponding final block for ResNet-50-based methods); (iii) feature maps are globally average-pooled along the channel dimension using the gradient magnitudes as weights; (iv) the resulting spatial map is ReLU-activated, upsampled to input resolution via bilinear interpolation, and min–max normalized to $[0, 1]$. All methods use the same Grad-CAM implementation and the same set of test images. Grad-CAM visualizations are qualitative: they indicate which spatial regions most influence a particular logit but do not provide localization guarantees.

D Scanpath Decoder Architecture

The input is resized to 448×448, yielding a 14×14 patch grid ($N = 196$) for finer spatial resolution. Patch features are projected from $d = 768$ to $d_s = 512$ via a linear layer. The decoder receives three inputs: the projected patch features, a learned task embedding $\mathbf{t} \in \mathbb{R}^{d_s}$ indicating which pathology to search for ($t \in \{0, \ldots, 12\}$), and the initial fixation coordinate $\mathbf{x}_0 = (0.5, 0.5)$, projected to

Table 7: Baseline comparison protocol. "Reported" indicates numbers taken from the cited paper under its original or unified evaluation protocol; "Rerun" indicates our own execution. For our method, the input resolution follows the corresponding evaluation setting: 224 for classification and 448 for scanpath probing, to match the protocol used in each experiment.

Method	**Backbone**	**Pretrain data**	**Resolution**	**Source**
BioViL	ResNet	MIMIC-CXR	224	Reported (EGMA)
MedKLIP	ResNet	MIMIC-CXR	224	Reported (EGMA)
MGCA	ResNet	MIMIC-CXR	224	Reported (EGMA)
GLoRIA	ResNet	CheXpert	224	Reported (EGMA)
PRIOR	ResNet	MIMIC-CXR	224	Reported (EGMA)
MedCLIP	Swin	MIMIC-CXR	224	Reported (EGMA)
CheXWorld	ViT	MIMIC-CXR	224	Rerun
EGMA	Swin	MIMIC-CXR + EYE	224	Reported (EGMA)
GazeFormer	ViT	ImageNet + COCO	224	Rerun
HAT	ResNet	ImageNet	224	Reported (GazeSearch)
GazeSearch	ResNet	ImageNet	448	Rerun
LogitGaze-Med	ResNet / CheX	MIMIC-CXR + EYE	448	Reported (LogitGaze)
Ours	Swin	MIMIC-CXR + EYE	224/448	Rerun

d_s dimensions. This centered initial fixation is supplied by the preprocessing protocol; it is not independently predicted.

At each step, the current fixation token queries the full set of patch features via cross-attention. A causal Transformer (6 layers, 8 heads, dimension $d_s = 512$) processes the growing fixation sequence with temporal positional encoding. Three output heads operate on each hidden state $\mathbf{h}_t$: (i) a spatial head, implemented as a 196-way softmax over the patch grid; (ii) a duration head that regresses fixation dwell time; and (iii) a termination head. Following the GazeSearch evaluation protocol, the decoder emits 7 fixations at inference time. The scanpath loss is

$$\mathcal{L}_{\text{scan}} = \mathcal{L}_{\text{spatial}}(\text{CE}, 196) + 0.1 \cdot \mathcal{L}_{\text{dur}}(\ell_1) + 0.1 \cdot \mathcal{L}_{\text{term}}(\text{BCE}). \tag{18}$$

Only the decoder parameters are trained during scanpath probing; the backbone remains frozen in this probing stage. The scanpath loss is never back-propagated into the backbone.

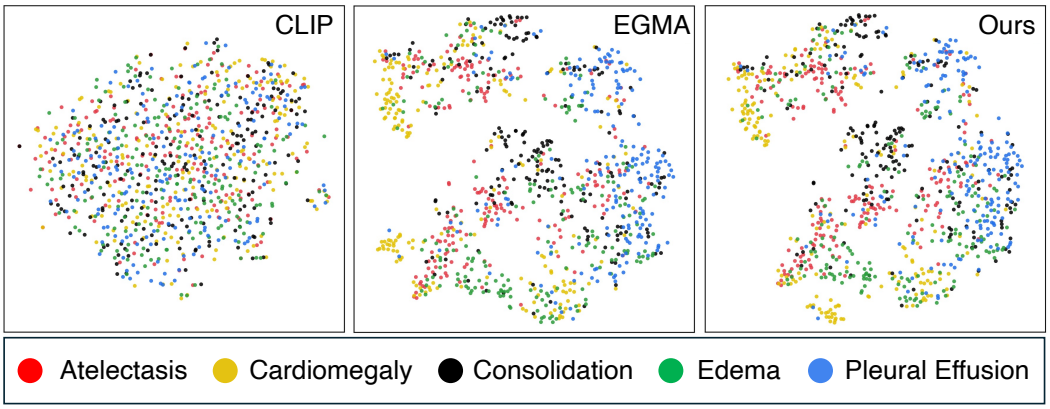


Figure 4: t-SNE visualization of learned representations on the CheXpert 5×200 dataset. Each point represents a chest radiograph and is color-coded by its ground-truth pathology label.

E Representation Visualization

Figure 4 provides a qualitative visualization of the learned representation space on the CheXpert 5×200 dataset. CLIP features show considerable mixing across pathology categories, while EGMA

produces partial separation with remaining inter-class overlap. Our method shows clearer pathology-level grouping, with more compact clusters for categories such as Cardiomegaly and Pleural Effusion. Since t-SNE [Van der Maaten and Hinton, 2008] is a qualitative projection and can be sensitive to hyperparameters, we use this visualization only as supporting evidence; the quantitative transfer results in Tables 1 and 2 provide the primary evaluation of representation quality.

F Additional Analyses and Experimental Results

This appendix provides method clarifications, gaze-free medical self-supervised baselines, initialization sensitivity, and resolution sensitivity. The controlled diagnostic order ablation is reported in the main text (Table 6). All AUROC values are percentages unless stated otherwise. Separate reruns and ablations are kept in separate tables; their point estimates are not pooled with the original main comparison.

F.1 What the Predictive Objective Learns

Behavioral supervision and image-derived targets. The observed expert scanpath is an external supervision signal. It determines which patch tokens are presented and in which temporal order. The latent target for a future fixation is computed from the actual image by the momentum target encoder, without a diagnostic label. Thus, "expert-gaze-guided predictive pretraining" distinguishes the source of the ordering supervision from the source of the prediction targets. Downstream classification still requires labels for its linear probe, and the scanpath decoder is explicitly trained on gaze annotations with a frozen backbone.

Predictable context, not unseen pathology. The predictor receives the prefix fixation tokens, while the next target token is used only in the loss. Prior observations need not determine the abnormality in a future region. Latent prediction instead encourages features to encode contextual dependencies among regions that experts attend to sequentially. Information unique to a future region need not be recoverable. Order ablations demonstrate the empirical usefulness of expert temporal structure under the evaluated encoding procedure; they do not establish a causal model of clinical reasoning or rule out all contextual shortcuts.

Temporal revisits and spatial discretization. Duplicate samples of the same patch within one time step are merged. A later return to that patch remains a separate fixation token, with its own temporal-rank and duration features. Thus, $\mathcal{S}$ may contain repeated patch indices, whereas $\mathcal{U}$ is the complement of the set of indices occurring anywhere in $\mathcal{S}$. Repeated visits and back-and-forth transitions across time steps are retained, but patch quantization still discards subpatch spatial precision. This clarification concerns the preprocessing definition; no separate revisit-removal ablation is reported.

Gaze context and completion queries. The contextual sequence $\mathbf{H} \in \mathbb{R}^{L\times 768}$ contains the predictor hidden states associated with the recorded fixation history. An unvisited location p supplies the query $\mathbf{m} + \mathbf{e}_p$, where $\mathbf{m}$ is the shared learnable mask token and $\mathbf{e}_p$ is its spatial embedding; $\mathbf{H}$ supplies keys and values. Table 8 gives the decoder configuration. The public implementation expands one parameter tensor of shape $1 \times 1 \times 768$ to all unvisited locations; only the learned spatial embeddings differ by position.

F.2 Comparison with Gaze-Free Medical Self-Supervised Models

Table 9 compares GazeWorld with CheXWorld [Yue et al., 2025], MedMAE [Gupta et al., 2025], and LVM-Med [Nguyen et al., 2023] using the same downstream linear-probing setting and 1%, 10%, and 100% labeled data. CheXWorld uses image world modeling, MedMAE uses masked reconstruction, and LVM-Med uses self-supervised graph matching. These models do not use expert gaze to define the prediction sequence. GazeWorld has the highest point estimate in all nine settings in this additional comparison. The results show the utility of combining an existing visual initialization with gaze-guided prediction; because pretraining data and initializations differ, this table alone does not isolate the causal effect of gaze supervision. The controlled order comparison in Table 6 more directly addresses that question.

Table 8: Spatial-completion decoder configuration. The target dimension d is 768 for the main MedCLIP-Swin configuration.

Component	Configuration
Depth	2 cross-attention blocks
Hidden dimension	768
Attention	12 heads, head dimension 64
Feed-forward network	SwiGLU, hidden dimension 2048
Normalization	RMSNorm
Attention projections	Bias-free query/key/value/output projections
Output head	Linear map from 768 to target feature dimension d
Query / keys and values	$\mathbf{m} + \mathbf{e}_p$ / contextual sequence $\mathbf{H}$

Table 9: Additional medical SSL comparison. Each entry lists AUROC at 1% / 10% / 100% labeled data. Values originally recorded on the $[0, 1]$ scale are converted to percentages without adding precision. These reruns are distinct from the point estimates in Table 1.

Method	CheXpert	RSNA	SIIM-ACR
GazeWorld	**78.8 / 83.5 / 87.6**	**83.2 / 85.6 / 90.1**	**88.0 / 90.4 / 94.2**
CheXWorld	65.7 / 81.4 / 85.5	78.7 / 84.1 / 88.5	68.0 / 73.9 / 91.3
MedMAE	62.2 / 69.1 / 76.6	74.9 / 77.6 / 85.8	77.7 / 82.7 / 90.2
LVM-Med (ResNet-50)	59.7 / 64.1 / 73.1	77.6 / 76.0 / 80.9	71.1 / 78.2 / 81.0
LVM-Med (ViT)	56.1 / 59.2 / 63.5	64.3 / 72.1 / 78.5	58.2 / 62.9 / 71.7

F.3 Sensitivity to Backbone Initialization

Table 10 adds medical, non-medical self-supervised, and random initializations. Each baseline and its GazeWorld-refined counterpart are evaluated with separate 100%-label linear probes on the three datasets. Positive gains occur for all five tested initializations, including DINOv3 [Siméoni et al., 2025] and random Swin. However, the random initialization remains substantially weaker in absolute AUROC, even after its larger numerical gains. MedCLIP gives the best final values among these settings. The evidence supports applicability to multiple tested encoders, while the magnitude of benefit depends on initialization and training configuration; it does not establish equal effectiveness for arbitrary backbones.

Table 10: Backbone sensitivity under 100%-label linear probing. Δ is the change in AUROC percentage points within each paired initialization experiment.

Initialization / refinement	CheXpert 5×200	RSNA	SIIM-ACR
MedCLIP-Swin	85.47	87.80	91.44
+ GazeWorld	**87.52**	**90.21**	**94.18**
Δ	+2.05	+2.41	+2.74
EGMA-Swin	85.45	87.96	89.05
+ GazeWorld	86.09	88.74	91.17
Δ	+0.64	+0.78	+2.12
DINOv3 ViT-L	75.08	84.81	89.22
+ GazeWorld	76.93	85.96	91.56
Δ	+1.85	+1.15	+2.34
CheXWorld ViT-B	85.62	88.54	91.30
+ GazeWorld	86.36	89.50	92.38
Δ	+0.74	+0.96	+1.08
Random Swin	61.32	73.75	63.46
+ GazeWorld	68.41	79.93	72.12
Δ	+7.09	+6.18	+8.66

F.4 Input Resolution

The main diagnostic experiments use 224×224 images to match the comparison setting. Scanpath probing uses 448×448 inputs as described in Appendix D. Table 11 reports an additional diagnostic run at 448×448, separate from the scanpath experiment. At 100% labels its AUROC is close to the main diagnostic point estimates. These results suggest limited gains on the evaluated benchmarks, but do not establish that 224×224 is sufficient for all clinically relevant findings.

Table 11: Diagnostic evaluation at 448×448: AUROC (%).

Labeled data	CheXpert	RSNA	SIIM-ACR
1%	79.11	83.40	88.12
10%	82.90	85.16	90.84
100%	87.91	90.43	94.44

F.5 Reader Variability and Extension Beyond Chest Radiographs

Reader and cohort coverage. GazeWorld uses existing expert eye-tracking records, with GazeSearch providing curated task-directed sequences derived from the same underlying data source. Curating gaze can reduce noise but does not establish representative coverage of reader expertise, viewing styles, institutions, or patient groups. We do not report reader-stratified training, held-out-reader evaluation, or a controlled comparison across experience levels. Therefore, external diagnostic transfer and positive gains across backbones cannot substitute for multi-reader or cross-institution gaze validation. An additional experiment that trains on one reader and tests transfer to another would directly address this limitation; such results are not available here.

Volumetric and dynamic imaging. The formulation could in principle treat fixated slices or 3D patches as sequential observations in CT or MRI, and use spatiotemporal observations for ultrasound. Such extensions would need 3D spatial and slice/time embeddings, synchronized gaze and navigation records, handling of slice scrolling and cross-slice revisits, and a completion objective appropriate for volumes or videos. Token counts and causal-attention cost would also grow. These are proposed adaptations, not evaluated capabilities of the present model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions and scope, including gaze-ordered world-model pretraining, diagnostic transfer, scanpath probing, and the evaluated chest radiograph setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 and Appendix F.5 discuss reader and domain bias, patch discretization, backbone dependence, statistical and localization limitations, and untested modalities.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results or formal proofs; the contribution is methodological and empirical.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The method and implementation settings are described in Section 3 and the appendices, and code is publicly available. Complete experiment records and a fully matched reproduction package are not provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is public at `https://github.com/levyisthebest/gazeworld` under the MIT license, with environment and training/probing instructions. Medical datasets require access under their original terms, and the release does not include every checkpoint and experiment record needed for full reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we have specific details in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The tables report point estimates. Confidence intervals, error bars, and significance tests are not included in this version, and no statistical significance is claimed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Architecture, epochs, batch sizes, and optimization settings are reported in Appendix B, but hardware, memory, runtime, and total compute are not fully documented.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses publicly available medical imaging and eye-tracking datasets under their intended research use and does not introduce procedures that conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Section 5 discusses potential label-efficiency benefits as well as reader/population bias, misplaced confidence in explanations, diagnostic error, and the need for independent clinical validation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk generative model, scraped dataset, or system intended for unrestricted clinical deployment; therefore, additional misuse safeguards are not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The original datasets, models, and baseline papers are credited, but the manuscript does not provide a complete asset-by-asset license and version inventory. Access to the source datasets remains governed by their original terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The implementation is released under the MIT license with a README, dependency requirements, and training/probing commands at `https://github.com/levyisthebest/gazeworld`. No new patient dataset is released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve new crowdsourcing experiments or new human-subject data collection; it uses existing publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The study does not involve new human-subject recruitment or data collection, and therefore does not raise new IRB-related concerns.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as part of the core method, experiments, or scientific contribution; they were used only for language polishing and formatting assistance.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.