X-WIN: Building Chest Radiograph World Model via Predictive Sensing

Supplementary Material

A. Datasets

Pretraining Datasets. We utilized public chest X-ray and CT datasets to pretrain X-WIN. We incorporated chest X-rays from MIMIC-CXR [8] containing 371,951 chest X-rays in frontal and lateral views from 65,086 patients. Chest X-rays in MIMIC-CXR characterize a wide range of abnormalities [9], ensuring the diversity of pretraining data. We incorporated chest CT scans from the NLST trial [4]. We utilized 32,371 CT scans from 9,353 patients in the NLST dataset for pretraining. We held out 10% of the NLST dataset (1,042 scans) as the unseen test set for model evaluation in the experiments.

Downstream Datasets. We evaluated model performance on six downstream chest X-ray datasets: VinDr [10], CheXpert [7], NIH-CXR [14], RSNA [1], JSRT [12], and COVIDx [11]. **VinDr** contains 15K images in the official training set and 3K images in the official test set. We evaluated multi-label classification performance via linear probing on 27 disease labels in the VinDr test set. **CheXpert** contains around 223K images in the official training set and 518 images in the official test set. We evaluated multi-label classification performance on five radiologist-annotated diseases in the CheXpert test set. **NIH-CXR** contains around 86K images in the training set and around 25K images in the test set. We evaluated multi-label classification performance on 14 diseases in the NIH-CXR test set. **RSNA** contains around 30K chest X-rays with binary labels of normal and pneumothorax. We split the dataset with a 70/15/15 ratio into train/val/test sets following [6]. We evaluated binary classification performance in the RSNA test set. **JSRT** contains 245 images with binary labels of normal and lung nodule. We split the dataset with a 70/10/20 ratio into train/val/test sets. We evaluated binary classification performance on the JSRT test set. **COVIDx** contains around 30K images in the official training set and 400 images in the official test set. COVIDx provides multi-class classification labels of normal, pneumonia, and COVID-19 pneumonia. We evaluated multi-class classification performance on the COVIDx test set.

B. Implementation Details

B.1. Pretraining

Cone-beam Projection Generation. Our generation of projections models the cone-beam geometry of a standard projectional radiography system. We utilized the DiffDRR algorithm [5] to generate cone-beam projections. In this approach, an X-ray source emits a ray through a CT volume and a pixel on the detector plane detects the accumulated attenuation experienced by the ray. DiffDRR employs a GPU-accelerated Siddon's method to compute plane intersections. In this work, we set the detector plane to contain 512×512 pixels and set a typical source-to-detector distance of 1,020 mm. We rotated the X-ray source with yaw rotation to obtain projections and restricted the rotation angle to the range of $[-90°, 90°]$ relative to the position of the context projection. Such a range captures sufficient spatial information for 3D reconstruction. We evaluated the computational efficiency of projection generation over 20 runs. The average time per generation is 0.219 ± 0.083 seconds.

Architectures. The X-WIN framework consists of four learnable networks: the encoder, view predictor, mask predictor, and domain classifier. We experiment with the encoder at different capacities using the ViT-L/16 and ViT-B/16 architectures with 303M and 86M parameters respectively. We utilize two separate networks for the view predictor and mask predictor. Following [2], we use a light-weight ViT architecture with 11M parameters for the predictors, containing 6 layers of transformer blocks with an embedding size of 384 and 12 heads for multi-head self-attention. The domain classifier is an attention pooling network with two transformer layers with an embedding size of 384 and 12 heads for multi-head self-attention and a global average pooling layer.

Masked Image Modeling. We utilized the multiblock masking strategy in [2] where the masked regions are the union of four random rectangular blocks within the scale $[0.15, 0.20]$. We concatenated the mask tokens with the unmasked context tokens produced by the encoder and added a sinusoidal positional embedding to construct the input to the mask predictor. The mask predictor reconstructs the masked tokens. We applied random resized cropping and color jitter as augmentations.

B.2. Experimental Details

Linear Probing and Fine-tuning. In linear probing experiments, we froze the exponential moving average encoder and trained a linear classifier on top to map the global average patch representations to classification logits. We used the AdamW optimizer to optimize the linear classifier and conducted a learning rate search in {1e-5, 2e-5, 5e-5, 1e-4, 2e-4, 5e-4, 1e-3, 2e-3, 5e-3} to obtain the best-performing learning rate. In fine-tuning experiments, we unfroze the encoder and set its learning rate to be one-tenth of the learning rate for the linear classifier. We conducted a learning rate search for best fine-tuning performance in the same

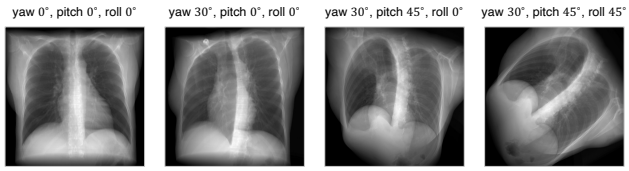

Figure 1. Example projections with three-dimensional rotation defined by yaw, pitch, and roll.

range as linear probing. For both linear probing and fine-tuning, we set the batch size to 64 for the VinDr, CheXpert, NIH-CXR, and RSNA datasets and 16 for the JSRT dataset.

Action Design. For the step-wise rotation, we defined the action as the rotation direction of clockwise/counterclockwise. We first obtained a routine X-ray projection u_{context} and then rotated the X-ray source clockwise/counterclockwise by $\Delta\phi$ to obtain the target projection u_{target}. We then conducted projection prediction with f_θ, $f_{\theta'}$, and g_v. At the following training iterations, we repeatedly treated the current u_{target} as the context and obtained the next projection as the target until we reached the target projection at $k \cdot \Delta\phi$. For the direct rotation (yaw, pitch, roll), the action is defined by yaw, pitch, and roll rotation angles. Specifically, to obtain a target X-ray projection, the X-ray source is rotated in the order of yaw, pitch, and roll to a defined angle. Example projections with three-dimensional rotation are shown in Fig. 1.

Effect of Domain Adaptation. To quantify representation similarity, we input images in the VinDr test set and the held-out NLST test set into $f_{\theta'}$ and obtained their average-pooled representations. We then obtained the cluster centers by computing the mean vectors of the groups of average-pooled representations from the VinDr test set and NLST test set, respectively. We quantified cosine similarity and L2 distance between the cluster centers of the two datasets.

Reconstruction of 3D CT Volumes. We trained a VQ-GAN decoder [3] on top of the frozen encoder f_θ and view predictor g_v to render 2D projections. Specifically, we first rearranged the g_v output $\mathbf{z}_i^{\text{patch}} \in \mathbb{R}^{L \times C}$ into 2D feature maps with a shape $H \times W \times C$, where L is the number of patch tokens, C is the feature dimension, and H and W denote spatial dimensions. To render the feature maps into projections, we conducted an element-wise quantization with a codebook and then inputted the quantized feature maps into a CNN decoder with consecutive convolutional and upsampling blocks. We set the decoder output resolution to 512×512 pixels and utilized ground-truth images to supervise rendered projections. We constrained the actions within $[-90°, 90°]$ with a step size $1°$. At inference time, we first obtained rendered projections within $[-90°, 90°]$ and then utilized the FDK algorithm in the ASTRA toolbox [13] to reconstruct a CT volume from these projections. We

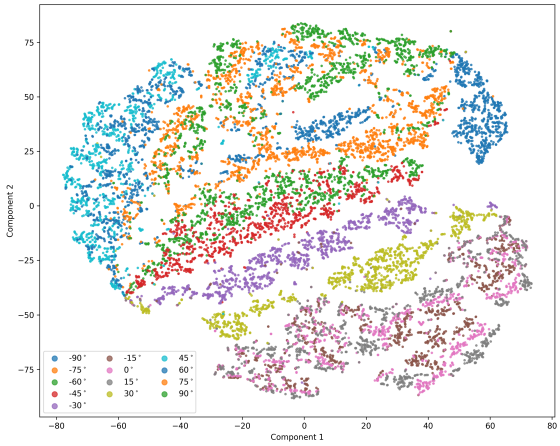

Figure 2. t-SNE visualization of the view predictor representations. X-WIN encodes representations that characterize distinct features for different projections while maintaining similarity between adjacent projections. The above representations were generated on the NLST held-out test set. Colors denote different actions.

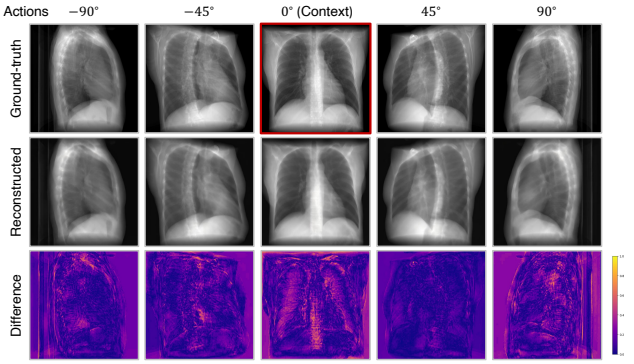

Figure 3. Visualizations of the rendered projections under different actions. Given a context projection and different projectional angles as actions, our world model encodes representations that capture the dynamics of structural changes.

visualized output representations conditioned on different actions in Fig. 2 and showed rendered projections in Fig. 3.

C. Additional Results

Fine-tuning Performance. We reported additional fine-tuning performance on the VinDr and RSNA datasets in Table 1. We sampled varied percentages of data from the training set to fine-tune the X-WIN encoder. Following [15], we conducted experiments with 6 global disease labels on the VinDr dataset. For the RSNA dataset, we evaluated the binary classification performance with labels of normal and pneumothorax. Results show that X-WIN attains improved performance over comparison CXR foundation models, demonstrating its effective adaptability and data efficiency.

Table 1. Model fine-tuning performance on additional datasets. We fine-tune the models with 1%, 10%, and 100% of samples in the training set. Five-run averages in AUROC are reported.

Models	VinDr			RSNA		
	1%	10%	100%	1%	10%	100%
CheXFound	0.856	0.871	0.893	0.847	0.869	0.883
RAD-DINO	0.789	0.815	0.837	0.831	0.858	0.876
Ark+	0.912	0.941	0.948	0.879	0.895	0.917
CheXWorld	0.898	0.936	0.951	0.815	0.841	0.854
X-WIN	**0.918**	**0.944**	**0.963**	**0.892**	**0.916**	**0.933**

References

[1] MD Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, MD Marc Kohli, Mark McDonald, Peter, Phil Culliton, Safwan Halabi MD, and Tian Xia. Rsna pneumonia detection challenge. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge, 2018. Kaggle. 1

[2] 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. 1

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[4] John Paul Flores, Alejandro Moreno-Koehler, Matthew Finkelman, Jaime Caro, and Gary Strauss. Ma03. 05 cost effectiveness analysis of ct vs chest x-ray (cxr) vs no screening for lung cancer (lc) in the plco and nlst randomized population trials (rpts). *Journal of Thoracic Oncology*, 12(1): S354–S355, 2017. 1

[5] Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Workshop on Clinical Image-Based Procedures*, pages 1–11. Springer, 2022. 1

[6] 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. 1

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 1

[8] 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. 1

[9] Mingquan Lin, Gregory Holste, Song Wang, Yiliang Zhou, Yishu Wei, Imon Banerjee, Pengyi Chen, Tianjie Dai, Yuexi Du, Nicha C Dvornek, et al. Cxr-lt 2024: A miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. *arXiv preprint arXiv:2506.07984*, 2025. 1

[10] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. 1

[11] Maya Pavlova, Tia Tuinstra, Hossein Aboutalebi, Andy Zhao, Hayden Gunraj, and Alexander Wong. Covidx cxr-3: A large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2206.03671*, 2022. 1

[12] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American journal of roentgenology*, 174 (1):71–74, 2000. 1

[13] Wim Van Aarle, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K Joost Batenburg, and Jan Sijbers. Fast and flexible x-ray tomography using the astra toolbox. *Optics express*, 24(22):25129–25147, 2016. 2

[14] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 1

[15] 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. 2