# CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning

Supplementary Material

## A. Datasets and Baselines

### A.1. Datasets

Our experiments are based on eight chest X-ray datasets, including MIMIC-CXR [13] for pre-training; CheXpert [11] and NIH ChestX-ray14 [26] for both pre-training and fine-tuning; VinDr-CXR [20], ShenZhen-CXR [12], RSNA Pneumonia [23], MedFMC-ChestDR [25], and SIIM-ACR Pneumothorax [1] for fine-tuning. Detailed information on these datasets is provided below.

- **MIMIC-CXR** [13] is one of the largest X-ray datasets, containing over 370k radiograph images from over 220,000 patient studies with paired radiology reports. We gather non-lateral scans from this dataset (about 230k images) and use this dataset for self-supervised pre-training.
- **CheXpert** [11] contains about 218k images with 14 disease labels automatically extracted from radiology reports. We use this dataset for pre-training and conduct multi-label classification experiments on five conditions: atelectasis, cardiomegaly, consolidation, edema, and effusion. We report the performance on the official validation set (200 patients) with a held-off subset from the training set for model selection. The mean AUROC score over the five classes is reported for this dataset.
- **NIH ChestX-ray14** [26] contains about 112k frontal-view chest radiographs, with annotations on 14 thoracic diseases: atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia, and pneumothorax. We use the training split of this dataset for pre-training and conduct disease classification experiments on the 14 classes. We follow the official split with 86k images for training and 25k for testing. The mean AUROC score over the 14 classes is reported for this dataset.
- **VinDr-CXR** contains 18,000 radiographs with expert annotations. Each radiograph is associated with 22 local findings and 6 global findings. We consider the multi-label classification task on the 6 global labels, including lung tumor, pneumonia, tuberculosis, other diseases, COPD, and no finding. We adopt the official split with 15,000 images for training and 3,000 images for testing. The mean AUROC score over the 6 classes is reported for this dataset.
- **ShenZhen-CXR** defines a binary classification problem where each radiograph is labeled with the presence of tuberculosis. We follow the data split provided by [18] with the train/val/test split containing 463/65/134 images, respectively. The AUROC score is reported for this dataset.
- **RSNA Pneumonia** [23] consists of over 26k radiographs, each categorized into one of three classes: normal, lung opacity, or no opacity but not normal. Additionally, expert-annotated bounding boxes highlight areas of lung opacity. This dataset is used for both classification and segmentation tasks. For classification, we frame it as a three-class problem, reporting top-1 accuracy. For segmentation, the bounding boxes are converted into segmentation masks and the mean dice score is reported. We follow the data split provided by [18] with train/val/test split containing 21295/2680/2709 images, respectively.
- **MedFMC-ChestDR** [25] is a dataset tailored for few-shot adaptation. Each radiograph is associated with 19 common thoracic disease labels. The official competition consists of 1-shot, 5-shot, and 10-shot tracks, each with five different train/val splits. To ensure consistency, we use the first split in each track and report the mean performance averaged over five random seeds. The mean AUROC score is reported over the 19 classes for this dataset.
- **SIIM-ACR Pneumothorax** [1] comprises 12,047 radiographs with pixel-level annotations for pneumothorax. We perform binary segmentation on this dataset, with the mean dice score reported as the evaluation metric.

### A.2. Baselines

We compare CheXWorld with several self-supervised learning methods developed for general-domain and medical images, including MoCo-v3 [6], DINO [5], BEiT [3], MAE [9, 27], SimMIM [18, 28], LVM-Med [19], Adam-v2 [10], and Rad-DINO [22]. When possible, we leverage radiology-specific adaptations of these methods. For a fair comparison, all methods utilize models of comparable sizes, such as ViT-B [7], Swin-B [15], and ConvNeXt-B [16]. Below, we provide a brief overview of each approach:

- **MoCo-v3** [6] is a contrastive learning framework that employs a momentum encoder to create a dynamic dictionary for stable and effective representation learning. It explores additional training techniques to optimize vision transformer performance.
- **DINO** [5] pre-trains vision transformers with a self-distillation objective. Techniques like distribution centering and sharpening are incorporated to stabilize the training process.
- **BEiT** [3] is a masked image modeling (MIM) approach inspired by masked language modeling in natural language processing. The model predicts masked token in-

dices generated by discrete variational autoencoders.

- **MAE** [9] is an encoder-decoder framework for MIM, predicting raw pixel values for masked patches. Only visible patches are passed to the encoder to improve computational efficiency. We use its radiology-adapted version introduced by [27].
- **SimMIM** [28] is another MIM approach based on the Swin Transformer [15]. It employs random masking with a moderately large patch size and uses a simple linear decoder head. The radiology-adapted version from [18] is used in our experiments.
- **LVM-Med** [19] leverages a graph-matching formulation for contrastive learning, building a versatile model that integrates diverse medical image modalities and datasets.
- **Adam-v2** [10] focuses on learning anatomical structures in X-ray images hierarchically, using pre-training objectives that promote localizability, composability, and decomposability.
- **Rad-DINO** [22] extends DINOv2 [21] by performing continuous pre-training on radiology datasets.

## B. Implementation Details

### B.1. Pre-training

**Data.** CheXWorld is pre-trained on the combination of three datasets: MIMIC-CXR [13], NIH ChestX-ray14 [26], and CheXpert [11] (following [27]). We only use the frontal scans for pre-training, resulting in ~0.5M radiographs in total. We exclude the validation/test split of the NIH Chest-Xray14 and CheXpert datasets from the pre-train dataset to avoid data leakage to the downstream tasks.

**Architecture and optimization.** The context encoder is a ViT-Base with a patch size of $16 \times 16$. The target encoder is the exponential moving average of the context encoder with an initial ratio equal to $0.996$ that gradually increases to $1.0$ following a cosine schedule. The predictor is 6 layers deep with 384-dimensional embeddings. We use sinusoidal functions [24] to encode the image patch positions following [9]. We use the AdamW optimizer [17] with $\beta_1 = 0.9, \beta_2 = 0.999$ with an initial learning rate of $2 \times 10^{-4}$ and weight decay set to $0.05$. Gradient clipping is set to $1.0$ throughout our experiments. The learning rate schedule follows linear warmup for 40 epochs and cosine annealing afterward. L2 loss is computed between the raw predictor outputs and the layer-normalized target encoder outputs. The model is trained from scratch for 300 epochs with a batch size of 2048, taking 16 hours on a machine with 8 RTX 4090 GPUs, each with 24 GB memory.

**Local anatomical structure modeling.** We adopt a block-wise masking strategy [2]. The image mask is the union of four rectangular blocks with the scale $(0.15, 0.2)$. We further shrink the context's visible area by a maximal factor of 0.25, which we found beneficial. The context encoder only processes unmasked patches, while the entire image takes the entire image as input. In the predictor, mask tokens corresponding to the masked locations are padded to the context. The loss is computed on masked locations.

**Global anatomical structure modeling.** We sample two random crops with the same spatial size with their scales in $(0.3, 1.0)$ and aspect ratios in $(0.75, 1.33)$. The relative position information $\Delta_{x\to y}$ is obtained in pixel space and then used to determine the location of target image patches in the context's coordinate system. Note that the sinusoidal encoding function $\mathrm{PE}(\cdot)$ supports fractional inputs. Thus, the target patch locations $\phi_{x\to y}(u, v)$ can be encoded in the same way as the context patch locations. We compute prediction loss on all target patches.

**Domain variation modeling.** We simulate domain transitions with a set of augmentations, including brightness, contrast, gamma transform, and Gaussian blur. Given an input image $I$ (or an image crop), the target is obtained by applying brightness and contrast adjustment to the original image. Then, we apply another augmentation consisting of bright, contrast, gamma transform, and Gaussian blur, with the configurations of the augmentation stored in the parameter $a$. In particular, $a$ consists of four scalars: the factor for brightness enhancement in the range $(0.6, 1.4)$, the factor for contrast adjustment in the range $(0.6, 1.4)$, the factor for gamma transform in the range $(0.5, 2.0)$ and the kernel size of the Gaussian blur in the range $(0.05, 2.0)$. Essentially, the context is obtained by augmenting the original image *twice*, where the second augmentation is modeled by CheXWorld. Domain variation modeling is implemented along with local or global anatomical modeling. The parameter $a$ is concatenated with the mask token $m \in \mathbb{R}^d$ along the feature dimension, resulting in a vector of length $d + 4$, which is then fed into the policy network $\pi$. The policy network $\pi$ is a three-layer MLP with an input dimension of $d + 4$ and an output dimension of $d$.

### B.2. Analytical Experiments

**Anatomical modeling visualization.** We utilize the RCDM framework [4] to showcase the anatomical modeling capabilities of CheXWorld. Specifically, we train a diffusion model to predict target pixel values conditioned on the output representation $\hat{h}_y$ of the world model. This guiding representation is first projected to a 512-dimensional vector, which is then injected into the diffusion model via conditional batch normalization layers [8] within each residual block. For local anatomical structure modeling, the diffusion model individually predicts four rectangular masked regions, guided by spatially pooled predictor outputs corresponding to each location. For global anatomical layout modeling, the model predicts the entire target region using spatially pooled outputs from the predictor. Figure 5 is built upon local anatomical modeling, focusing on masked regions with visible artifacts. The diffusion model is trained using the validation split of the NIH ChestXray-14 dataset, while the visualizations are generated from the test split. This separation ensures that there is no information leakage between the different stages of the experiment—CheXWorld pre-training, diffusion model training, and visualization.

**Anatomical Correspondence Visualization.** We input the entire radiograph into the CheXWorld encoder to obtain image patch embeddings. Then we calculate per-pixel feature embeddings using RoI pooling over a 2x2 window centered on the pixel location. To illustrate anatomical correspondence, we focus on four key anatomical landmarks: the aortic arch, right hilum, left ventricle, and clavicle. The final similarity map is computed by measuring the L2 distance between the landmark embeddings of the reference image and the pixel embeddings of the test image. For improved visualization, the similarity values are rescaled.

**Domain sensitivity test.** To evaluate how effectively CheXWorld handles domain variations, we construct a test dataset using different augmentation configurations applied to the same base image. Specifically, we sample $n = 64$ augmentation parameters evenly from a predefined range and apply these augmentations to generate a candidate set of target images $\{y_i\}_{i=1}^n$. For each target $y_i$, we further apply a randomly sampled augmentation to obtain the corresponding context $x_i = \mathcal{T}_{a_i}(y_i)$, resulting in a set of context-target-latent triplets $\{(x_i, y_i, a_i)\}$. The model's task is to predict the target $y_i$ given the context $x_i$ and latent $a_i$. The prediction error is defined as:

$$L(y; x, a) = \|g(f_\theta(x); a) - f'_{\theta'}(y)\|^2. \tag{1}$$

Ideally, the prediction error $L(y_i, x_i, a_i)$ should be smaller than $L(y_j, x_i, a_i)$ for any $j \neq i$. For the $i$-th case, we rank the errors $\{L(y_j, x_i, a_i)\}_{j=1}^n$ across the candidate set and compute the top-k recall rate of the true target $y_i$. This procedure is repeated across 50 different images, and the final result is the averaged outcome over these trials.

### B.3. Fine-tuning

For classification, we employ mean pooling over all the output tokens to obtain a global feature representation of the image. Subsequently, a task-specific linear head is attached to the model for fine-tuning. We utilize the AdamW optimizer with a default learning rate of $1 \times 10^{-4}$, with layer-wise decay set to $0.75$ and a drop path rate of $0.6$. For the CheXpert benchmark, we adopt a learning rate of $1 \times 10^{-2}$ and a drop path of $0.1$. The data augmentation pipeline involves random resized cropping and color jittering.

For segmentation, we connect a U-Net decoder with the pre-trained backbone with a SimpleFPN [14] adapter. The U-Net decoder has four stages with number of channels 8, 16, 32, and 64. The initial learning rate is set to $2 \times 10^{-4}$ with a layer-wise decay rate of $0.8$ and a drop path rate of $0.1$. The data preprocessing pipeline for training involves random brightness contrast, shifting, and scaling.

Due to the varying sizes of the datasets, we employ different batch sizes and epochs across benchmarks. The input size of the image is set to $224 \times 224$ pixels. 10% of the training data is used for validation. Each experiment is conducted five times.

## C. Numerical Results

Figure 7 illustrates the fine-tuning performance of CheXWorld on the VinDr-CXR dataset using varying proportions of the training data, which highlights CheXWorld's ability to enhance data efficiency. Here, we present the corresponding numerical results in Table 1.

Table 1. Fine-tuning with 1%, 10%, and 100% training data on VinDr-CXR.

| Method | 1% | 10% | 100% |
|---|---|---|---|
| LVM-Med | 76.41±3.79 | 85.85±0.59 | 88.22±0.44 |
| Adam-v2 | 77.90±1.14 | 88.26±0.48 | 91.46±0.33 |
| MAE | 78.07±1.66 | 90.63±0.16 | 92.76±0.18 |
| SimMIM | 83.85±1.62 | 92.15±0.31 | 92.81±0.31 |
| CheXWorld | **90.53±1.01** | **94.71±0.10** | **95.24±0.12** |

## References

[1] Zawacki Anna, Wu Carol, Shih George, Elliott Julia, Fomitchev Mikhail, Hussain Mohannad, Lakhani Paras, Culliton Phil, and Bao Shunxing. Siim-acr pneumothorax segmentation. 2019. 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. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[4] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2

[10] Mohammad Reza Hosseinzadeh Taher, Michael Gotway, and Jianming Liang. Representing part-whole hierarchies in foundation models by learning localizability, composability, and decomposability from anatomy via self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 1, 2

[11] 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, 2

[12] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6): 475, 2014. 1

[13] 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, 2

[14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 3

[15] 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. 1, 2

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1

[17] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 2

[18] DongAo Ma, Mohammad Reza Hosseinzadeh Taher, Jiaxuan Pang, Nahid UI Islam, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Benchmarking and boosting transformers for medical image classification. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 12–22. Springer, 2022. 1, 2

[19] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[20] 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

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 2

[22] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024. 1, 2

[23] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 1

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[25] Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023. 1

[26] 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, 2

[27] 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. 1, 2

[28] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 1, 2