This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning

Yang Yue[1*] Yulin Wang[1*] Chenxin Tao[1] Pan Liu[2] Shiji Song[1] Gao Huang[1✉]
[1]Tsinghua University [2]PLA General Hospital
yueyang22@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn

Abstract

Humans can develop internal world models that encode common sense knowledge, telling them how the world works and predicting the consequences of their actions. This concept has emerged as a promising direction for establishing general-purpose machine-learning models in recent preliminary works, e.g., for visual representation learning. In this paper, we present CheXWorld, the first effort towards a self-supervised world model for radiographic images. Specifically, our work develops a unified framework that simultaneously models three aspects of medical knowledge essential for qualified radiologists, including 1) local anatomical structures describing the fine-grained characteristics of local tissues (e.g., architectures, shapes, and textures); 2) global anatomical layouts describing the global organization of the human body (e.g., layouts of organs and skeletons); and 3) domain variations that encourage CheXWorld to model the transitions across different appearance domains of radiographs (e.g., varying clarity, contrast, and exposure caused by collecting radiographs from different hospitals, devices, or patients). Empirically, we design tailored qualitative and quantitative analyses, revealing that CheXWorld successfully captures these three dimensions of medical knowledge. Furthermore, transfer learning experiments across eight medical image classification and segmentation benchmarks showcase that CheXWorld significantly outperforms existing SSL methods and large-scale medical foundation models. Code & pre-trained models are available at https://github.com/LeapLabTHU/CheXWorld.

1. Introduction

Intelligent agents like humans learn extensive background knowledge about the world [39]. This common-sense information is embedded in the agents' internal models of the world, playing a pivotal role in their perception, learning, and decision-making processes by simulating the world's dynamics, telling them what is plausible or impossible, and predicting the outcomes of their actions. Consequently, these world models enable agents to acquire new concepts and skills with minimal demonstrations and trials [55]. Recent works [2, 21] have preliminarily verified the effectiveness of establishing visual representation learning approaches through the lens of world modeling. Pre-trained world models can produce semantically rich embeddings and adapt to various downstream tasks with limited data.

Like many other research areas, the field of medical imaging is experiencing a paradigm shift from task-specific models to general-purpose foundation models [20, 34, 45, 68], which are pre-trained on massive data and expected to encode meaningful medical knowledge. The idea of world modeling offers an approach to training a medical foundation model by capturing common sense information (*e.g.* human anatomy) from medical images, which is a promising yet under-explored research direction.

In this paper, we present CheXWorld, the first initiative toward world modeling for self-supervised representation learning on radiographic images. We propose and integrate three world modeling tasks that capture different dimensions of medical knowledge essential for qualified radiologists, depicted in Figure 1. In particular, a radiologist with a strong understanding of human anatomy can identify different anatomical structures and determine the relative geometry between anatomical regions. Given the hierarchical nature of human anatomy, we introduce world modeling tasks that focus on both local and global levels of anatomy: *local anatomical structure modeling* aims to predict specific anatomical characteristics and structures within a localized anatomical region, such as bones, airways, blood vessels, and lung segments; *global anatomical layout modeling* learns the overall arrangement and spatial relationships of various anatomical regions within the human body, understanding the relative positioning of organs and tissues such as the heart, lungs, diaphragm, and rib cage. Additionally, medical images are

*Equal contribution. ✉Corresponding authors.

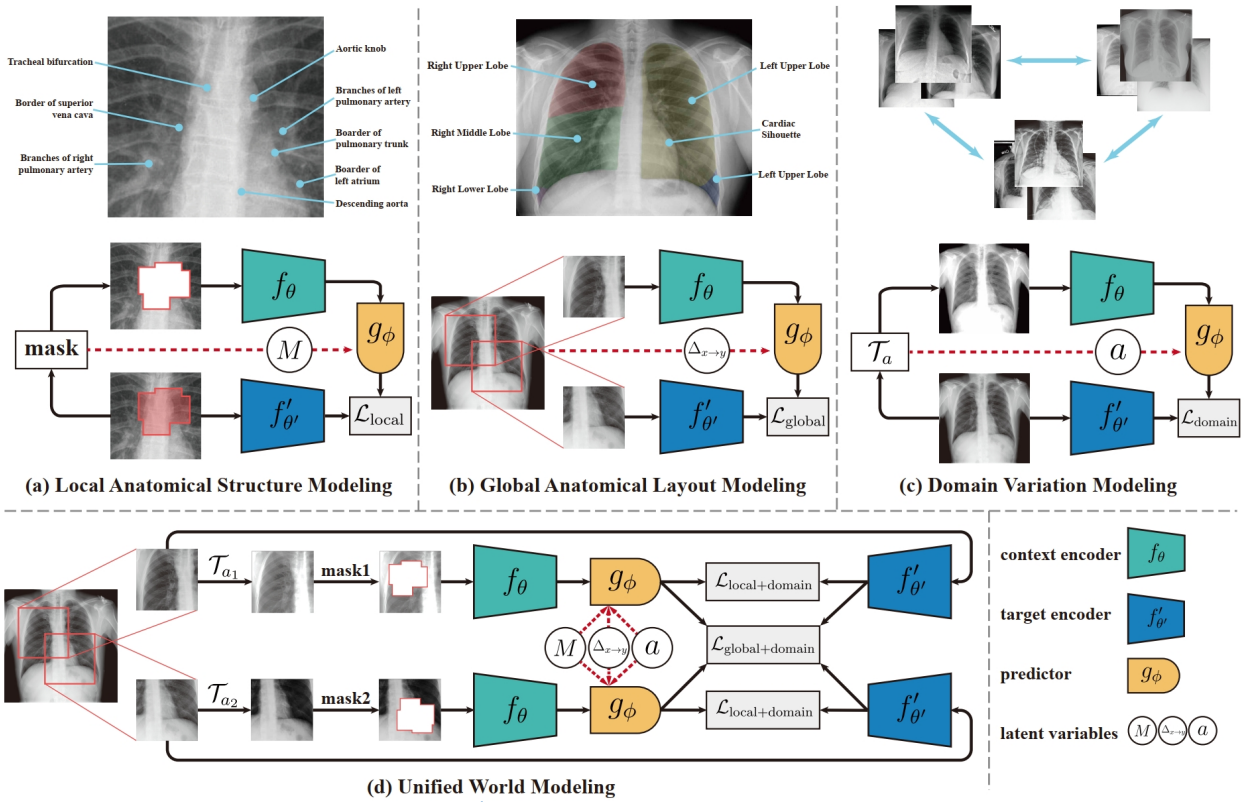


Figure 1. **Overview of the CheXWorld framework**[1]**.** The upper part of the figure depicts three dimensions of medical knowledge that are formulated in our framework, including (a) local anatomical structures describing the fine-grained characteristics of local tissues, (b) global anatomical layouts describing the global organization of the human body and (c) domain variations that encourage CheXWorld to model the transitions across different appearance domains of radiographs. The middle part of the figure illustrates the world modeling tasks corresponding to these aspects of medical knowledge. (d) shows our unified pipeline that combines the merits of all three tasks.

typically gathered from multiple domains with notable appearance differences due to varying equipment and acquisition techniques [52]. Experienced radiologists can recognize the physical reality behind domains with different appearances and facilitate the diagnosis process. Motivated by this, we hypothesize that radiologists quickly adapt to diverse domain changes by learning sophisticated internal world models that are capable of simulating transitions between domains. Driven by our hypothesis, we introduce the *domain variation modeling* task, which aims to learn an expressive feature space that encodes predictable transitions between domains. Finally, we design an integrated pipeline that performs the three world modeling tasks simultaneously, effectively incorporating medical anatomical knowledge and achieving robustness across domains.

It is noteworthy that the three aspects we consider, *i.e.*, local anatomical structure modeling, global anatomical layout modeling, and domain variation modeling, represent core elements of radiographic expertise, each contributing to a comprehensive framework for capturing medical knowledge. While they may not encompass every dimension of medical understanding, they are fundamental to radiology and critical for radiograph representation learning, as extensively demonstrated in [23, 49, 69, 70].

Empirically, we validate that CheXWorld effectively captures the three dimensions of medical knowledge through a series of analytical experiments, including visualizations of predictor outputs with the help of a generative model and a domain variation sensitivity test. Moreover, we extensively evaluate CheXWorld on eight medical image classification and segmentation benchmarks. Our model consistently outperforms competitive self-supervised learning baselines with comparable backbone capacity, pre-training data, and pre-training computational cost.

2. Related Work

World Modeling. In the machine learning community, world modeling originates from model-predictive control [9, 10] and is a common practice in reinforcement learning, where it predicts the future state of the environment based on the agent's action [24–27]. Recently, building general world models featuring profound comprehension of common sense knowledge is believed to be a crucial step towards general artificial intelligence [39, 53]. Video gener-

[1]The x-rays are taken from kenhub.com, Radiopedia [43] and ChestX-ray14 [61].

ation models [37] trained on vast data (*e.g.* SoRA[47] and VideoPoet [38]) are world models that perform predictions in input space. Another line of work makes predictions in the latent space [2, 7, 28, 33, 39], which more closely resembles the world model in the human brain. World modeling is also an effective approach for visual representation learning, where the model predicts missing parts of the visual inputs [2, 4, 5, 7, 21]. Our framework builds on this foundation, where our model acquires knowledge—such as human anatomy and domain variations—through learning to predict unobserved outcomes from specific actions. Our key contribution lies in providing insights into modeling medical knowledge from radiographic images and designing a framework that unifies three world modeling tasks.

Self-supervised learning in medical imaging. Self-supervised learning (SSL) [19, 31, 48, 62] provides a practical solution to mitigating annotation scarcity in medical imaging. A large body of works incorporate restorative SSL tasks, which typically involve reconstructing the original data from certain image corruptions, such as random masking [63, 67, 68], pixel shuffling [69], and patch order shuffling [50, 58, 70]. Another line of works [3, 12, 51, 56, 66] adapt global or dense contrastive learning to medical imaging. TransVW[29] and DiRA [30, 57] explore combining discriminative and restorative methods. Additionally, learning from anatomy is a common topic for medical SSL. SAM [65] enforces pixel-level anatomical correspondence between images, and Adam [32] learns part-whole correspondence that reflects the hierarchy of human anatomy. Recently, SSL-based foundation models trained on large-scale datasets gained significant attention, such as RETFound [68] for retinal imaging and UNI [13] for pathology.

Compared with existing works, CheXWorld is the first to introduce the concept of world modeling into medical SSL, capturing task-relevant medical knowledge through three tailored world modeling tasks. Unlike contrastive-based approaches such as Adam [32], CheXWorld takes a technically distinct path by constructing equivariant [16, 17] representations—where input transformations result in predictable changes within the embedding space—rather than invariant representations. This approach allows CheXWorld to better model medical data without discarding valuable information. Moreover, our unified SSL framework delivers state-of-the-art performance across eight benchmarks, demonstrating the significant potential of designing SSL methods grounded in the philosophy of world modeling.

3. Basic Framework of World Modeling

World modeling. We begin by briefly introducing the concept of *world modeling* [25, 39], which forms the basis for our proposed method. The primary motivation behind *world modeling* is to predict the unobserved parts y of a world (*e.g.*, a visual environment, an image, or a video) based on an observed context x. This prediction can be formulated across various dimensions, such as spatial (predicting the unseen regions of the data) and temporal (foreseeing the consequence of an action). In fact, the paradigm of *world modeling* takes inspiration from human visual cognition. For example, when presented with the right upper lobe region of a chest radiograph as the context x, the internal world of a radiologist could imagine that there are two lobes (right middle and right lower) below the visible region and two lobes on the left , forming the target y.

A basic framework. Here, we introduce our basic framework of instantiating world modeling, upon which we will further discuss how to establish a radiology world model by formulating various medical knowledge in Section 4. In specific, our basic framework mainly follows the joint-embedding predictive architecture (JEPA) [2, 39], as shown in Figure 2. We consider two encoders: the *context encoder* f_θ and the *target encoder* $f'_{\theta'}$, which encode the context x and target y into representations h_x and h_y, respectively. A *predictor* g_ϕ learns to predict the target embedding h_y using the context embedding h_x, conditioned on an additional latent variable z that indicates how the target relates to the context. For example, z can be the spatial locations of y with respect to x, or an action that leads to the transition from context to target $x \rightarrow y$ (*e.g.*, image transformations). The latent variable z carries the information that makes y predictable based on x, effectively modeling the inherent variation and uncertainty of the real world. The primary objective of world modeling is:

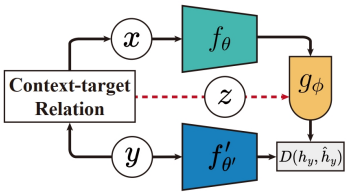


Figure 2. **A basic framework of world modeling.**

$$\text{minimize} \quad D(h_y, \hat{h}_y) = D(h_y, g_\phi(h_x; z)), \tag{1}$$

where $D(\cdot, \cdot)$ denotes the prediction error. Notably, the predictions are performed in the abstract representation space, contrasting with generative world models that predict every detail of the target. This mirrors an important characteristic of intelligent agents: information filtering, *i.e.*, eliminating the irrelevant details in the data during the perception process. For example, a radiologist can infer the anatomical layout of a partly missed image, but it's nearly impossible for him to recover each pixel value of the unseen regions.

In implementation, we set the target encoder to be the exponential moving average of the context encoder following [2, 4, 7]. Moreover, we use the Vision Transformer (ViT) [18] architecture for the context/target encoder and the predictor. For the rest of this paper, the context and target feature h_x, h_y are **sequences of patch embeddings** produced by the ViTs. To perform patch-level feature predictions, we attach mask tokens to the context features and feed the com-

bined sequence to the predictor. The output representations of the mask tokens are used as the final prediction.

4. CheXWorld: World Modeling for Radiograph Representation Learning

This section extends the idea of world modeling to radiology. We propose three world modeling tasks tailored for radiographs and establish a unified CheXWorld framework that seamlessly integrates them. Importantly, the three tasks are designed to model three critical dimensions of medical knowledge essential for qualified radiologists. To be specific, these tasks are organized hierarchically from low-level to high-level, namely 1) *local anatomical structure modeling* that aims to learn the fine-grained structures of local anatomical regions, 2) *global anatomical layout modeling* that seeks to learn the global geometry of the human body (*e.g.*, the layout of organs and skeletons), and 3) *domain variation modeling* that learns to model the transition across different appearance domains of radiographs. We first elaborate on the details of each task. Then, we present a unified framework that comprehensively combines the characteristics and merits of all three tasks, yielding a powerful foundation model that produces semantically rich and transferable representations.

4.1. Local Anatomical Structure Modeling

At a relatively low, fundamental level, various types of tissues (e.g., bones, muscles, and epithelial tissues) form the structural organization of the human body. Understanding the fine-grained characteristics (*e.g.*, shapes, sizes, appearance, and textures) of these micro-structures is essential knowledge that radiologists must possess. To build up a learning procedure that enables neural networks to acquire such local anatomical knowledge, we consider a mask-and-reconstruction task. The model predicts fine-grained details of a masked tissue region based on its peripheral, surrounding information, as shown in Figure 1(a). This encourages the networks to develop an internal model that captures the intricate structures and local continuities of human microanatomy. The large, continuous prediction target also prevents reliance on low-level features.

Specifically, the image mask is a union of four randomly selected rectangular regions of the image following [2]. Let M denote the set of 2D patch locations that are masked from the context input. Given an input image (crop) I, the image patches with locations contained in M are dropped from the image to create the context $x = \mathrm{Mask}(I, M)$, while the target is the entire image. The context and target are then fed to the corresponding encoders, producing representations $h_x = f_\theta(x)$ and $h_y = f'_{\theta'}(y)$. Note that the context encoder exclusively processes visible patches, following [2, 31]. In the predictor, the mask tokens m carrying the positional encoding of the masked locations are concatenated with the context feature h_x, which forms the input token sequence to the predictor. The predictor output $\hat{h}_y$ is given by:

$$\begin{aligned}\hat{h}_y &= g_\phi(h_x; M) \\ &= g_\phi\left((h_x + p_x) \oplus \{m + \mathrm{PE}(u, v)\}_{(u,v)\in M}\right),\end{aligned} \tag{2}$$

where $\oplus$ is the concatenation operation along sequence dimension, p_x is the positional embedding corresponding to the context, u and v are image patch location indices along height and width dimensions, $\mathrm{PE}(\cdot)$ is the sinusoidal positional encoding proposed by [59]. The prediction loss is computed only on the masked locations, which is given by:

$$\mathcal{L}_{\mathrm{local}}(x, y) = \sum_{c\in M} \|g_\phi(f_\theta(x); M)_c - f'_{\theta'}(y)_c\|_2^2, \tag{3}$$

where c is the 2D image patch location that belongs to M.

4.2. Global Anatomical Layout Modeling

In addition to understanding local microstructures like tissues, it is also important for radiologists to possess comprehensive medical knowledge regarding how groups of tissues or systems of organs are anatomically organized and assembled into the full bodies of humans. Inspired by this, we develop the global anatomical layout modeling task (see Figure 1(b)), seeking to predict the feature of an out-of-context area based on its relative position to a given context, thus facilitating understanding of the global structure of human bodies. Notably, this idea is orthogonal to the technique proposed in section 4.1 (see section 5.3 for ablation studies) as it focuses on formulating the long-range topological relationships of different tissues or organs instead of the fine-grained anatomical structures within local tissues.

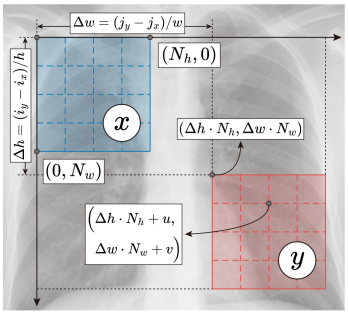


Figure 3. **Formulation of global anatomical layout modeling.**

As illustrated in Figure 1(b), we randomly crop two anatomical areas from the original radiograph, serving as context x and target y, respectively. The world model learns to predict y from x given their relative position information $\Delta_{x\to y}$. In particular, we assume the context and target image crop share the same spatial size of h pixels in height and w pixels in width, with the left-top corner located at (i_x, j_x) and (i_y, j_y) in the pixel space. The relative position information is defined as the proportional relative displacement between the two regions:

$$\Delta_{x\to y} = (\Delta h, \Delta w) = \left(\frac{i_y - i_x}{h}, \frac{j_y - j_x}{w}\right). \tag{4}$$

To properly represent the location of the image patches in the target y, we establish a coordinate system over the context image patches, with the top-left corner patch located at the origin $(0, 0)$ and the bottom-right corner patch located at $(N_h - 1, N_w - 1)$, where N_h, N_w are the number of tokens along the height and width dimensions. As demonstrated in Figure 3, the coordinates of the image patch at the u-th row and v-th column in the target image are given by:

$$\begin{aligned} \phi_{x\to y}(u, v) &= (\Delta h \cdot N_h + u, \Delta w \cdot N_w + v) \\ &\text{where} \quad u \in [0, N_h - 1], v \in [0, N_w - 1]. \end{aligned} \tag{5}$$

Essentially, $\phi_{x\to y}(u, v)$ describes the target's location from the context's perspective, enabling the model to predict the target based on the context. Sinusoidal position encoding $\mathrm{PE}(\cdot)$ is then applied to the relative coordinates $\phi_{x\to y}(u, v)$, and the mask tokens carrying the positional embeddings of all the target image patches are fed to the predictor. The predictor output is given by:

$$\begin{aligned} \hat{h}_y &= g_\phi(h_x; \Delta_{x\to y}) \\ &= g_\phi\left((h_x + p_x) \oplus \{m + \mathrm{PE}(\phi_{x\to y}(u, v))\}_{u,v}\right), \\ &\text{where} \quad u \in [0, N_h - 1], v \in [0, N_w - 1], \end{aligned} \tag{6}$$

and the loss function is given by:

$$\mathcal{L}_{\text{global}}(x, y) = \|g_\phi(f_\theta(x); \Delta_{x\to y}) - f'_{\theta'}(y)\|_2^2. \tag{7}$$

4.3. Domain Variation Modeling

Thus far, we have discussed two levels of anatomical knowledge of human bodies. Beyond this, radiologists are usually capable of flexibly adapting to the specific characteristics of each radiograph and understanding them despite variations [15, 23, 49]. As shown in Figure 1(c), medical images often come from different sources (*e.g.*, hospitals, techniques, equipment) and show diverse appearances [52] (*e.g.*, clarity, contrast, exposure). Nevertheless, experienced radiologists can imagine the physical reality behind the radiographs with varying appearances, which facilitates the diagnosis process. We hypothesize that radiologists' internal world models can simulate the variations across domains, thus obtaining an objective view of the scanned body.

Inspired by this, we propose a domain variation modeling task that learns how image features change across domains, enabling cross-domain adaptability. We simulate domain shifts using data augmentation and construct context-target pairs to model these transitions. The model is tasked with predicting the feature outcomes of the inverse effect of an image transformation determined by certain augmentation parameters, as shown in Figure 1(c). The learned representation space is designed to be equivariant [16, 17], *i.e.*, input transformation leads to predictable output change, which preserves sufficient information across domains, mimicking radiologists' internal models.

Specifically, the target y is an augmented version of the original image. Another transform parameterized by k scalars (denoted by $a \in \mathbb{R}^k$) is further applied to the target to produce the context $x = \mathcal{T}_a(y)$, where $\mathcal{T}$ is an augmentation pipeline consisting of Gaussian blur, brightness, contrast, and gamma adjustments. The parameter a contains the strength and other configurations of the augmentation. The world model learns to model the feature transformation from the context to the target domain (*i.e.* the inverse effect of $\mathcal{T}_a$) conditioned on the parameter a, which can be regarded as the "action" of domain transition. The prediction is also performed with mask tokens to be compatible with the other two world modeling tasks. The parameter a is combined with the mask tokens via a lightweight policy network π. The forward pass and loss function of domain variation modeling are given by:

$$\begin{aligned} m_{a,p} &= \pi(m + p, a), \\ \hat{h}_y &= g_\phi(h_x; a) \\ &= g_\phi\left((h_x + p_x) \oplus \{m_{a,\mathrm{PE}(u,v)}\}_{u,v}\right), \\ \mathcal{L}_{\text{domain}}(x, y) &= \|g_\phi(f_\theta(x); a) - f'_{\theta'}(y)\|_2^2, \end{aligned} \tag{8}$$

where p is the positional embedding of an image patch, $m_{a,p}$ denotes the mask token carrying both spatial and domain transition information.

4.4. The Unified Framework

Having established the individual tasks of local anatomical structure modeling, global anatomical layout modeling, and domain variation modeling, we now integrate these tasks into a cohesive system. This integration is important for creating a comprehensive model that can leverage multiple dimensions of medical knowledge simultaneously and maximize the combined benefits of each task.

Instead of simply adding three loss functions together, a more elegant solution is to design unified contexts, targets, and latent variables to perform the three world modeling tasks in a single forward pass. As illustrated in Figure 1(d), the training pipeline of CheXWorld starts with sampling two different image crops from the image y_1, y_2, which serve as the targets. The targets are then augmented and masked with two different configurations (a_1, M_1) and (a_2, M_2), producing the contexts x_1, x_2:

$$x_i = \mathrm{Mask}(\mathcal{T}_{a_i}(y_i), M_i),\ i = 1, 2. \tag{9}$$

The contexts x_1, x_2 and targets y_1, y_2 are then encoded into representations $h_{x_1}, h_{x_2}, h_{y_1}, h_{y_2}$. Note that the targets h_{y_1}, h_{y_2} are predictable from the contexts h_{x_1}, h_{x_2} given the proper conditions. Specifically, since x_1 is obtained by applying augmentation $\mathcal{T}_{a_1}$ and mask M_1 to y_1, predicting y_1 from x_1 with conditions (a_1, M_1) simultaneously performs local anatomical structure modeling and domain vari-

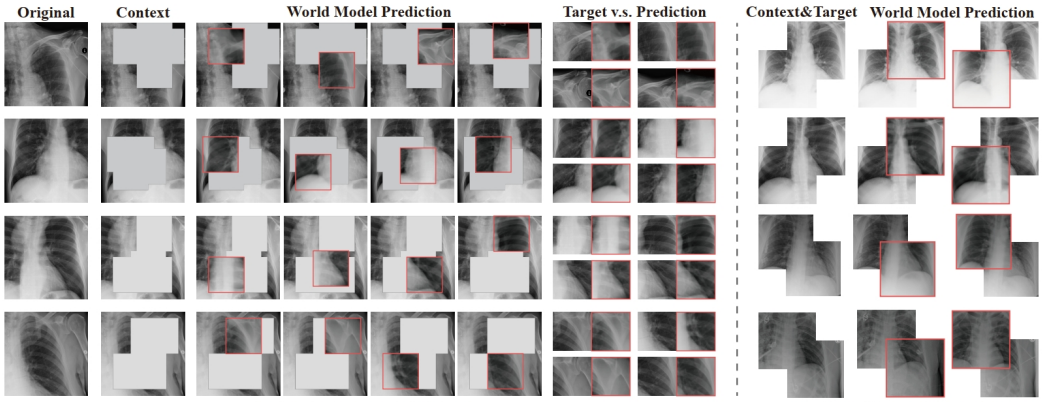


Figure 4. **Visualization of the CheXWorld predictor outputs** (zooming in for details). The images presented in this figure were not included in the pre-training of CheXWorld or the training of the diffusion model. Regions in red bounding boxes denote the predictor outputs that are mapped to pixel space using the RCDM [8] framework. In (a), gray areas indicate masked regions excluded from the context. In (b), the two overlapping regions alternately serve as context and target.

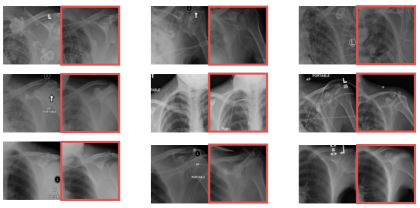


Figure 5. **CheXWorld prioritizes relevant medical features over spurious signals** (*e.g.*, lateral markers) in the image. Regions in red boxes denote the predictor outputs.

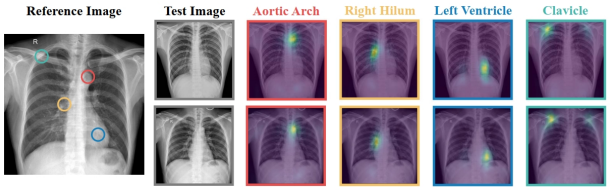


Figure 6. **CheXWorld learns anatomical correspondence.** We compute pixel-level embeddings using RoI-pooling [22] and calculate the embedding similarity between four anatomical landmarks in a reference image and each pixel in the test image to create feature similarity heatmaps.

ation modeling:

$$\begin{aligned}\mathcal{L}_{1\to1} &= \mathcal{L}_{\text{local+domain}}(x_1, y_1) \\ &= \sum_{c\in M_1} \|g_\phi(h_{x_1}; M_1, a_1)_c - f'_{\theta'}(y_1)_c\|_2^2 .\end{aligned} \tag{10}$$

Moreover, x_1 can be derived from y_2 by moving the input window by $-\Delta_{x_1\to y_2}$ and applying augmentation $\mathcal{T}_{a_1}$, hence the predictor conditioning on $(\Delta_{x_1\to y_2}, a_1)$ performs global anatomical structure modeling and domain variation modeling at the same time.

$$\begin{aligned}\mathcal{L}_{1\to2} &= \mathcal{L}_{\text{global+domain}}(x_1, y_2) \\ &= \|g_\phi(h_{x_1}; \Delta_{x_1\to y_2}, a_1) - f'_{\theta'}(y_2)\|_2^2 .\end{aligned} \tag{11}$$

Similarly, $\mathcal{L}_{2\to2}, \mathcal{L}_{2\to1}$ can be computed between (x_2, y_2) and (x_2, y_1), respectively. Note that the context features and target features are reused twice when computing the loss (10) and (11). Finally, the model is trained with the combination of $2 \times 2 = 4$ supervisions computed from the two pairs of context and target.

$$\mathcal{L} = \mathcal{L}_{1\to1} + \mathcal{L}_{2\to2} + \mathcal{L}_{1\to2} + \mathcal{L}_{2\to1}. \tag{12}$$

With this unified objective, CheXWorld showcases deep comprehension of human anatomy in different scales and pixel intensity distributions across various domains.

5. Experiments

In this section, we first demonstrate CheXWorld's world modeling capability with a series of analytical experiments. Then, we empirically compare CheXWorld with other self-supervised methods as baselines on eight downstream tasks. Finally, we provide ablation studies to highlight the effect of each building block of our method.

Implementation details. CheXWorld is pre-trained on ∼0.5M frontal chest X-rays from several public datasets. The visual backbone is a ViT-Base [18]. The model is trained for 300 epochs, taking 16 hours on 8 RTX 4090 GPUs. Please refer to the appendix for more details.

5.1. CheXWorld is a Strong World Predictor

Visualization of anatomical modeling. To better understand the predictions made by CheXWorld, we train a generative model that maps the average-pooled predictor output back to pixel space following the RCDM framework [8]. Specifically, we construct context-target pairs in the same way as in the local and global anatomical modeling tasks. Then, we feed the context to a frozen pre-trained CheXWorld model to obtain the model's feature prediction, on which a diffusion model conditions to produce the target image in pixel space. As shown in Figure 4(a), the CheXWorld predictor perfectly recovers the appearance of bones

Table 1. **Performance on five downstream classification benchmarks.** Accuracy is reported for RSNA, and the area under the ROC curve (AUROC) is reported for the rest of the datasets. The best two results are **bold-faced** and underlined, respectively.

Method	Pre-train Data	Backbone	VinDr-CXR	ShenZhen	ChestX-ray14	RSNA (CLS)	CheXpert
Scratch	-	ViT-B	70.22±1.95	82.24±0.60	71.69±0.32	66.59±0.39	80.78±0.03
MoCo-v3 [14]	IN-1k (1.3M)	ViT-B	87.25±0.63	92.85±1.00	79.20±0.30	72.79±0.52	87.12±0.36
DINO [11]	IN-1k (1.3M)	ViT-B	82.89±1.10	90.39±4.29	78.37±0.47	71.27±0.45	87.01±0.62
BEiT [6]	IN-21k (12M)	ViT-B	85.93±1.98	92.87±1.08	79.91±0.24	72.78±0.37	87.77±0.38
LVM-Med [44]	Medical (1.3M)	ViT-B	88.22±0.44	94.81±1.32	80.08±0.09	72.75±0.44	88.07±0.25
MAE [31, 63]	X-rays (0.5M)	ViT-B	92.76±0.18	97.63±0.21	83.0†	73.75±0.24	89.3†
SimMIM [42, 64]	IN-21k (12M) & X-rays (0.9M)	Swin-B	92.81±0.31	98.09±0.13	83.04±0.15	74.09±0.39	89.14±0.22
Adam-v2 [32]	IN-21k (12M) & X-rays (0.9M)	ConvNeXt-B	91.46±0.33	97.80†	83.4†	73.40±0.88	88.90±0.36
CheXWorld	X-rays (0.5M)	ViT-B	**95.24±0.13**	**98.88±0.06**	**83.58±0.05**	**75.03±0.39**	**89.63±0.13**
Rad-DINO‡ [51]	LVD (142M) & X-rays (public+private)	ViT-B	95.16±0.16	98.20±0.17	83.61±0.08	74.51±0.46	88.94±0.15

Results on ImageNet pre-trained models are adopted from [42]. †Results reported by the original authors [31, 32]. ‡Rad-DINO [51] is included for reference but not for direct comparison, as it requires 2560 GPU-hours (20 times more than ours) and is trained on both public and private datasets.

and tissues within various masked areas, and Figure 4(b) shows that the model makes global-level predictions that are consistent with the context. Further visualizations also show that CheXWorld filters artifacts in the image (Figure 5) and learns anatomical correspondence (Figure 6).

Awareness to domain variations. We design a domain sensitivity test to verify the model's ability to capture the uncertainty in a radiograph's appearance attributes. Specifically, we first generate a candidate set of targets $\{y_i\}_{i=1}^n$ by applying augmentations with different configurations to the same image. Then, we construct a set of context-target-latent triplets $\{(x_i, y_i, a_i)\}$ by applying another augmentation $x_i = \mathcal{T}_{a_i}(y_i)$. The model is asked to predict the target y_i given the context x_i and latent a_i. For each predicted output $\hat{y}_i$, we calculate the top-k recall rate of the true target y_i over the entire candidate set using L2 distance[2]. As shown in Table 4, our model achieves an average top-5 recall of 77.67 (10 times higher than random choice), demonstrating the model's strong discriminative ability across domains. We also observe a significant drop in the recall rate when removing the domain condition a, indicating that the model indeed makes predictions conditioned on a.

5.2. CheXWorld Produces Transferable Representations

Setup and baselines. We compare CheXWorld with several competitive baselines across eight classification and segmentation tasks using the following datasets: VinDr-CXR [46], ShenZhen-CXR [36], NIH ChestX-ray14 [61], CheXpert [35], MedFMC-ChestDR [60], SIIM-ACR Pneumothorax [1] and RSNA Pneumonia [54].

[2]In our experiments the candidate set size is fixed at $n = 64$. This process is repeated across multiple images, and the final result is obtained by averaging the outcomes.

Table 4. **Effectiveness of domain variation modeling.**

Method	Recall@1	Recall@3	Recall@5
Random Choice	1.56	4.68	7.81
w/o domain condition	15.87	33.12	44.37
CheXWorld	**38.49**	**65.90**	**77.67**

We attach task-specific heads (linear heads and U-Net decoders) and perform full fine-tuning on the pre-trained CheXWorld encoder. The input image resolution is fixed to 224^2. We compare CheXWorld with several self-supervised learning methods designed for general domain images [6, 11, 14, 31, 64] and medical images [32, 44] using base-scale backbones [18, 40, 41]. We use the radiology version of these methods [42, 51, 63] if available.

Classification results are shown in Table 1. Our method consistently outperforms the best alternatives on all the benchmarks. Although SimMIM and Adam-v2 are initialized from ImageNet weights and pre-trained on a large dataset (0.9M X-rays) and, yet still lag behind CheXWorld with considerable margins. For instance, CheXWorld surpasses Adam-v2 by 4.04 in AUROC on VinDr-CXR and SimMIM by 0.54 on ChestX-ray14. Our approach performs comparably to Rad-DINO across all benchmarks, despite Rad-DINO requiring 20 times more computational resources and utilizing private training data. Additionally, one can observe a significant contrast in performance between models trained on photometric images (ImageNet) and models trained on medical images, highlighting the importance of in-domain transfer.

Segmentation results are presented in Table 2 and Figure 8. CheXWorld surpasses the leading baselines in terms of dice score and produces more accurate segmentation masks, demonstrating its superior capability in both instance-level and dense prediction settings.

Few-shot learning results on the MedFMC dataset are presented in Table 2. CheXWorld demonstrates a clear ad-

Table 2. **Results on segmentation (left) and few-shot learning (right) tasks.** The dice score and the AUROC score are reported for the segmentation and few-shot learning benchmarks respectively.

Method	Segmentation		Few-shot (MedFMC-ChestDR)		
	SIIM-ACR	RSNA (SEG)	1-shot	5-shot	10-shot
LVM-Med [44]	82.19±0.30	78.47±0.24	57.55±0.81	66.95±0.60	67.44±0.71
MAE [31, 63]	83.01±0.39	77.39±0.25	61.31±0.27	74.03±0.30	75.26±0.36
SimMIM [42, 64]	82.89±0.07	78.31±0.41	60.64±1.92	72.14±0.54	74.42±0.38
Adam-v2 [32]	83.60±0.33	78.53±0.19	59.16±1.03	70.28±0.39	70.67±0.71
CheXWorld	**84.58±0.34**	**79.02±0.25**	**64.60±1.00**	**75.19±0.51**	**76.40±0.25**

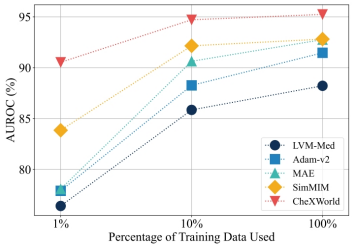


Figure 7. Fine-tuning with 1%, 10%, and 100% training data on VinDr-CXR.

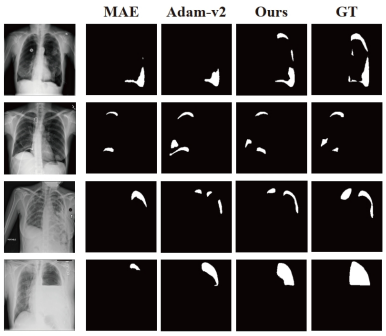


Figure 8. **Visualization of segmentation masks** on SIIM-ACR pneumothorax dataset. GT stands for the ground truth masks.

Table 3. **Ablation studies of the world modeling tasks and latent variables.** We report the results on VinDr-CXR and RSNA with 1% proportion of training data. AUROC and accuracy are reported for the respective datasets. "Local.", "Global.", and "Domain." are abbreviations for local anatomical structure modeling, global anatomical layout modeling, and domain variation modeling, respectively. "Unified" indicates whether our unified formulation is applied, as opposed to a naive combination of losses.

World Modeling Task			Latent Variables		Unified	VinDr 1%	RSNA 1%
Local.	Global.	Domain.	$\Delta_{x\to y}$	a			
✓					N.A.	84.71±1.00	65.89±0.54
✓	✓		✓		✓	88.31±0.88	67.47±0.51
✓		✓		✓	✓	89.73±0.50	68.27±1.00
✓	✓	✓	✓	✓	✓	**90.53±1.01**	**70.01±0.33**
✓	✓	✓	✓	✓		89.12±0.76	68.37±0.50
✓	✓	✓	✓		✓	89.91±0.60	68.98±0.60
✓	✓	✓		✓	✓	87.60±0.72	66.56±0.39

vantage on the 1-shot learning task and consistently outperforms the second-best methods in both 5-shot and 10-shot scenarios, showcasing quick adaptation to new concepts by leveraging knowledge acquired through world modeling.

Sample efficiency results.[3] The superiority of our method is further highlighted by reducing the training data proportion. As shown in Figure 7, we conduct experiments on VinDr-CXR with 1%, 10%, and 100% of the training data. Notably, our method with 10% of the data significantly outperforms all baselines trained on the entire dataset, reducing the annotation cost by over 90 percent.

5.3. Ablation Study

Effectiveness of the world modeling tasks is validated in the upper part of Table 3. Combining local anatomical structure modeling with either of the other two world modeling tasks yields significant performance gains. The best performance is achieved when all three world modeling tasks are integrated within our unified framework, indicating that these tasks reflect different aspects of medical knowledge that are effectively captured by CheXWorld.

Role of the latent variables. To verify that the latent variables (such as relative position $\Delta_{x\to y}$ for global anatomical layout modeling and transformation parameter a for domain variation modeling) contribute to transfer learning performance, we test the models pre-trained without certain latent variables, shown in the bottom part of Table 3. Domain variation modeling without the condition a degrades into a blind image restoration task, which yields inferior results compared with the original method. Removing the condition $\Delta_{x\to y}$ from global anatomical layout modeling leads to a sharp decrease in performance because it confounds the model with erroneous target spatial locations.

[3] Detailed numerical results can be found in the appendix.

6. Conclusion

In this paper, we proposed CheXWorld, a self-supervised world modeling framework for radiographic images that simultaneously encodes local and global anatomical knowledge and appearance domain variations. By integrating three world modeling tasks into a unified pipeline, our model effectively captures the spatial and domain uncertainties of the image world, as verified by qualitative and quantitative analytical experiments. Extensive transfer learning experiments on eight medical image analysis benchmarks demonstrate the state-of-the-art performance of our model. Our work offers new insights into representing and extracting knowledge from medical images, paving the way toward a general-purpose foundation model for medical vision.

Acknowledgements.

The work is supported in part by the National Key R&D Program of China under Grant 2024YFB4708200 and National Natural Science Foundation of China under Grant 62321005.

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. 7

[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, 4

[3] 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. 3

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 3

[5] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023. 3

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 7

[7] 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. 3

[8] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 6

[9] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. Routledge, 2018. 2

[10] Eduardo F Camacho, Carlos Bordons, Eduardo F Camacho, and Carlos Bordons. *Constrained model predictive control*. Springer, 2007. 2

[11] 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. 7

[12] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in neural information processing systems*, 33:12546–12558, 2020. 3

[13] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 3

[14] 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. 7

[15] Peng Cui, Yang Yue, Zhijie Deng, and Jun Zhu. Confidence-based reliable learning under dual noises. *Advances in Neural Information Processing Systems*, 35:35116–35129, 2022. 5

[16] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021. 3, 5

[17] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations*, 2023. 3, 5

[18] 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. 3, 6, 7

[19] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[20] Cong Gao, Benjamin D Killeen, Yicheng Hu, Robert B Grupp, Russell H Taylor, Mehran Armand, and Mathias Unberath. Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308, 2023. 1

[21] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024. 1, 3

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 6

[23] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 2, 5

[24] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018. 2

[25] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 3

[26] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[27] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 2

[28] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3

[29] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021. 3

[30] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022. 3

[31] 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. 3, 4, 7, 8

[32] 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. 3, 7, 8

[33] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 3

[34] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 1

[35] 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. 7

[36] 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. 7

[37] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 3

[38] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[39] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022. 1, 2, 3

[40] 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. 7

[41] 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. 7

[42] 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. 7, 8

[43] Phillip Marsh. Case courtesy of phillip marsh, radiopaedia.org. https://radiopaedia.org/cases/58938, 2023. Case ID: rID: 58938. 2

[44] 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. 7, 8

[45] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 1

[46] 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. 7

[47] OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. Accessed: Feb, 2024. 3

[48] 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. 3

[49] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. 2, 5

[50] Jiaxuan Pang, Fatemeh Haghighi, DongAo Ma, Nahid Ul Islam, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Popar: Patch order prediction and appearance recovery for self-supervised medical image analysis. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 77–87. Springer, 2022. 3

[51] 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. 3, 7

[52] Eric Postal. Why don't radiology textbooks have imperfect images? https://www.diagnosticimaging.com/view/why-dont-radiology-textbooks-have-imperfect-images, 2019. Accessed: Feb, 2024. 2, 5

[53] RunwayML. Introducing general world models. https://research.runwayml.com/introducing-general-world-models, 2023. Accessed: May, 2023. 2

[54] 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. 7

[55] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119, 2022. 1

[56] 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. 3

[57] Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Caid: a self-supervised learning framework for empowering instance discrimination in medical imaging. *Proceedings of Machine Learning Research*, 3, 2022. 3

[58] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 238–248. Springer, 2020. 3

[59] 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. 4

[60] 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. 7

[61] 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. 2, 7

[62] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5852–5864, 2023. 3

[63] 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. 3, 7, 8

[64] 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. 7, 8

[65] Ke Yan, Jinzheng Cai, Dakai Jin, Shun Miao, Dazhou Guo, Adam P Harrison, Youbao Tang, Jing Xiao, Jingjing Lu, and Le Lu. Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. *IEEE Transactions on Medical Imaging*, 41(10):2658–2669, 2022. 3

[66] Hong-Yu Zhou, Shuang Yu, Cheng Bian, Yifan Hu, Kai Ma, and Yefeng Zheng. Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 398–407. Springer, 2020. 3

[67] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–6. IEEE, 2023. 3

[68] 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. 1, 3

[69] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 2, 3

[70] Ziyu Zhou, Haozhe Luo, Jiaxuan Pang, Xiaowei Ding, Michael Gotway, and Jianming Liang. Learning anatomically consistent embedding for chest radiography. *arXiv preprint arXiv:2312.00335*, 2023. 2, 3