Leveraging the Structure of Medical Data for Improved Representation Learning

Andrea Agostini [*1] **Sonia Laguna** [*1] **Alain Ryser** [*1] **Samuel Ruiperez-Campillo** [*1] **Moritz Vandenhirtz** [1]
Nicolas Deperrois [2] **Farhad Nooralahzadeh** [2] **Michael Krauthammer** [2] **Thomas M. Sutter** [†1] **Julia E. Vogt** [†1]

Abstract

Building generalizable medical AI systems requires pretraining strategies that are data-efficient and domain-aware. Unlike internet-scale corpora, clinical datasets such as MIMIC-CXR offer limited image counts and scarce annotations, but exhibit rich internal structure through multi-view imaging. We propose a self-supervised framework that leverages the inherent structure of medical datasets. Specifically, we treat paired chest X-rays (i.e., frontal and lateral views) as natural positive pairs, learning to reconstruct each view from sparse patches while aligning their latent embeddings. Our method requires no textual supervision and produces informative representations. Evaluated on MIMIC-CXR, we show strong performance compared to supervised objectives and baselines being trained without leveraging structure. This work provides a lightweight, modality-agnostic blueprint for domain-specific pretraining where data is structured but scarce.

1. Introduction

Advances in foundation models have ignited interest in clinical AI, yet the data realities of medicine differ sharply from the internet-available corpora that power vision-language behemoths. Within the medical domain, in the popular field of radiology, chest-X-ray collections such as MIMIC-CXR comprise hundreds of thousands—not billions—of images, and every expert annotation carries substantial monetary and time cost (Hassanzadeh et al., 2018). Consequently, frontier general-purpose models still lag behind experts on domain benchmarks and remain costly to fine-tune or deploy in practice (Chaves et al., 2024). To learn informative representations without incurring those expensive medical labels, self-supervised learning has emerged as a useful approach (Tiu et al., 2022; Azizi et al., 2021). Among the many approaches, domain-adapted masked autoencoders, which reconstruct held-out patches, and complementary contrastive objectives that distinguish latent views have proven effective for distilling rich features from unlabeled exams (Xiao et al., 2023). Importantly, these techniques are most effective when the pre-text task, visible-patch ratio, and augmentations are re-engineered around medical imagery's limited scale and anatomical regularities (Mo & Liang, 2024).

Beyond clever objectives, the latent organization of clinical datasets is an under-explored source of supervision. Medical studies frequently present multi-view pairs, longitudinal follow-ups, and paired image–report examples, all of which encode consistent anatomy or semantics across views. Exploiting such coherence, whether through multi-view contrast, cross-modal alignment, or multitask training, has been shown to improve robustness and label efficiency in recent multimodal radiology models (Mo & Liang, 2024; Pellegrini et al., 2025; Chen et al., 2024). Therefore, capitalizing on these built-in redundancies offers a scalable path toward domain-specific foundation models.

The present work follows this intuition, proposing a multi-view regularized masked autoencoder framework that leverages paired chest-X-ray views to learn view-invariant representations without relying on external supervision. Unlike recent vision–language models that depend on paired reports (Pellegrini et al., 2025; Chen et al., 2024), our method exploits the complementary information provided by multiple X-ray views acquired during a patient examination. Concretely, we treat the frontal–lateral views in each MIMIC-CXR exam as natural positives. For every view, we run a masked autoencoder that reconstructs its own missing patches, regularizing the latent space. In a separate head we apply a regularization objective, e.g. a contrastive loss that pulls together the frontal and lateral embeddings of the same study while pushing apart embeddings from different samples. Optimized jointly, these complementary objectives yield features that are simultaneously detail-rich and view-invariant, with no textual supervision or annotation burdens of image–report pairs. Although we focus on chest radiographs, the same strategy could extend to any life-science

[*] Equal contribution [†] Joint senior authorship [1] Department of Computer Science, ETH Zurich, Zurich, Switzerland [2] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. Correspondence to: <{anandrea, slaguna, aryser, sruiperez}@inf.ethz.ch>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

domain that offers repeated or complementary scans (e.g., longitudinal MRIs or multi-sequence CT).

Our study makes two principal contributions: (i) We show that explicitly exploiting the internal structure of clinical datasets enables effective pretraining on limited data, outperforming supervised training on downstream tasks; (ii) we design the first multi-view MAE and contrastive pipeline for radiology and demonstrate consistent gains on MIMIC-CXR generalizing across modalities, establishing a lightweight, modality-agnostic blueprint for future medical foundation models.

2. Dataset and Preprocessing

MIMIC-CXR as a bimodal resource. We conduct all experiments on the publicly available *MIMIC-CXR* archive (Johnson et al., 2019b), a large-scale collection of routine chest radiographs acquired in critical-care settings. Each imaging *study* bundles every projection that shares one radiology report and fourteen diagnostic labels derived by CheXpert (Irvin et al., 2019). Image quality varies markedly because of patient positioning, bedside hardware, and emergent clinical constraints (Raoof et al., 2012). To expose inherent view redundancy, we isolate two complementary projection families—*frontal* (posterior–anterior *PA* or anterior–posterior *AP*) and *lateral* (left-lateral *LL* or generic *Lateral*). Whenever a study contains at least one image from each family, we enumerate every frontal–lateral combination to create paired samples (see Figure 1), yielding a collection $\mathcal{X} = \{\mathbf{X}^{(i)}\}_{i=1}^{N}$ with $\mathbf{X}^{(i)} = \{\mathbf{x}_f^{(i)}, \mathbf{x}_l^{(i)}\}$, where $\mathbf{x}_f^{(i)}$ and $\mathbf{x}_l^{(i)}$ are a frontal and a lateral radiograph from the i^{th} study and $\mathcal{V} = \{f, l\}$ defines the set of views. A raw image may appear in multiple tuples, yet every pair is unique by construction. This framing supplies natural positive pairs for self-supervision while preserving the study-centric semantics of the original dataset.

Preprocessing pipeline and data split. Each radiograph is center-cropped and isotropically resized to 224×224. Intensities are rescaled to $[0, 1]$ and standardized with ImageNet statistics. Study-level labels are inherited from the MIMIC-CXR-JPG release (Johnson et al., 2019a), produced by the CheXpert labeler. Following Haque et al. (2023), the three non-positive states ("negative", "not mentioned", "uncertain") are collapsed into a single 0-class, treating only explicit positives as 1. To prevent patient leakage, we partition *subjects*—and thus all associated studies and image pairs—into training (80%), validation (10%), and test (10%) splits.

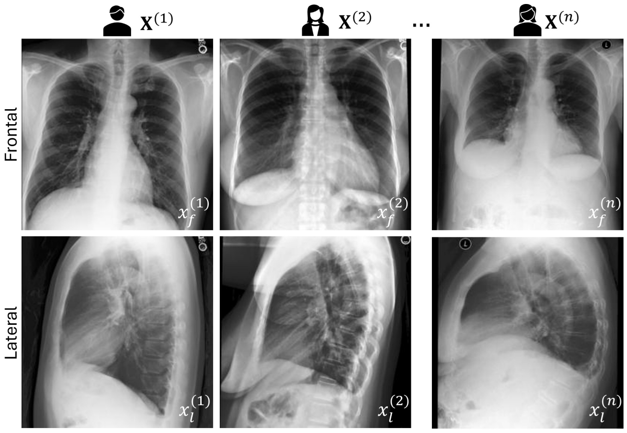

Figure 1. Frontal–lateral pairs from MIMIC-CXR. Each column shows a frontal image ($\mathbf{x}_f^{(i)}$, top) and its matching lateral view ($\mathbf{x}_l^{(i)}$, bottom).

3. Method

We compare two different pretraining paradigms in this work. The first method combines a reconstruction loss with an additional alignment loss between views, where the second approach applies a multi-view contrastive learning approach (Tian et al., 2020). An overview of the two pretraining paradigms is shown in Figure 2.

Multiview MAE MAEs (He et al., 2022) are a self-supervised learning approach designed to learn high-quality representations by reconstructing missing portions of the input. MAEs randomly mask a large fraction of the input data and train an encoder-decoder architecture to recover the masked content from the visible subset. In this work, we assume the encoder and decoder to be vision transformers (ViTs, Dosovitskiy et al., 2020).

The encoder f_θ processes only the unmasked input tokens $\mathbf{x}_{\text{vis}}^{(i)} = (1 - M(\mathbf{x}^{(i)}))$ producing latent representations $\mathbf{z}^{(i)} = f_\theta(x_{\text{vis}}^{(i)})$. $M(\cdot)$ applies a random mask to the input to mask. These are passed, along with mask token placeholders, to a lightweight decoder g_ϕ, which attempts to reconstruct the original input x, including the masked parts. The model is trained by minimizing a reconstruction loss $\mathcal{L}_{\text{Rec}}$ over only the masked positions:

$$\mathcal{L}_{\text{Rec}}(\mathbf{x}^{(i)}) \propto \sum_{t \in M(\mathbf{x}^{(i)})} \left\| \mathbf{x}_t^{(i)} - g_\phi(f_\theta(\mathbf{x}_{\text{vis}}^{(i)}))_t \right\|_2^2 .$$

Here, $\mathbf{x}_t^{(i)}$ denotes the original input at masked position t, and $g_\phi(f_\theta(\mathbf{x}_{\text{vis}}^{(i)}))_t$ is the decoder's reconstruction at that position with $\mathbf{x}_{\text{vis}}^{(i)}$ being the visible or un-masked part of the input image. For more details, we refer to He et al. (2022).

We propose a late fusion approach to adapt the MAE ob-

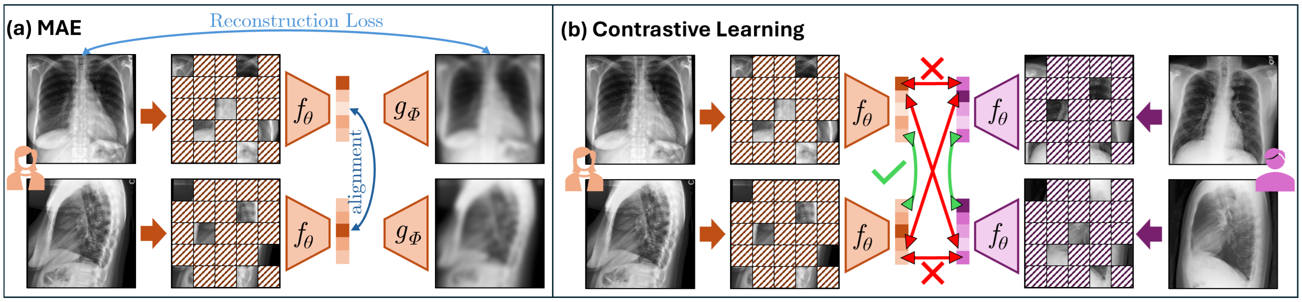

Figure 2. Pretraining strategies for multi-view chest radiographs. **(a)** *MAEs* reconstruct masked patches from visible ones using an encoder–decoder architecture and optionally enforce alignment between frontal and lateral views. **(b)** *Contrastive Learning* maximizes agreement between corresponding views in the same study while contrasting against other samples in the batch using a contrastive loss.

jective to the multi-view setting. In addition to the reconstruction, we emphasize similarity between the latent representations through an additional alignment objective $\mathcal{L}_{\text{Align}}$ (Sutter et al., 2024; Agostini et al., 2024).

$$\mathcal{L}_{\text{Align}}(\mathbf{x}_f^{(i)}, \mathbf{x}_l^{(i)}) \propto \sum_{t=1}^{T} d_{\text{MSE}}(f_\theta(\mathbf{x}_f^{(i)})_t, f_\theta(\mathbf{x}_l^{(i)})_t), \quad (1)$$

where $d_{\text{MSE}}(\cdot,\cdot)$ is the mean squared error (MSE). In this work, we only consider the MSE as an alignment metric, but the extension to other measures is part of future work.

The objective of the proposed multi-view MAE approach, MVMAE, for a pair of images x_f and x_l follows as

$$\mathcal{L}(\mathbf{x}_f^{(i)}, \mathbf{x}_l^{(i)}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{Rec}}(\mathbf{x}_m^{(i)}) + \beta \cdot \mathcal{L}_{\text{Align}}(\mathbf{x}_f^{(i)}, \mathbf{x}_l^{(i)}),$$

where β is an additional weighting between reconstruction and alignment loss.

Contrastive Learning To encourage view-invariant representations across paired frontal and lateral chest radiographs, we adopt a contrastive learning objective inspired by SimCLR (Chen et al., 2020). For each image pair $\mathbf{X}^{(i)} = \{\mathbf{x}_f^{(i)}, \mathbf{x}_l^{(i)}\}$ in a study, we encode both views using a shared transformer encoder backbone and extract latent representations from the [CLS] token, as used in Radford et al. (2021). The contrastive loss is computed as:

$$\mathcal{L}_{\text{Con}}(\mathcal{X}) = -\frac{1}{2N} \sum_{v \in \mathcal{V}} \sum_{i=1}^{N} \Gamma(\mathcal{X}, v, i), \quad (2)$$

where

$$\Gamma(\mathcal{X}, v, i) = \log\left(\frac{\Lambda(\mathbf{x}_v^{(i)}, \mathbf{x}_{\bar{v}}^{(i)})}{\sum_{\substack{k=1 \\ k \neq i}}^{N} \Lambda(\mathbf{x}_v^{(i)}, \mathbf{x}_v^{(k)}) + \Lambda(\mathbf{x}_v^{(i)}, \mathbf{x}_{\bar{v}}^{(k)})}\right),$$

where $\Lambda(\mathbf{x}_v^{(i)}, \mathbf{x}_{\bar{v}}^{(i)}) = \exp\left(\text{sim}(f_\theta(\mathbf{x}_v^{(i)}), f_\theta(\mathbf{x}_{\bar{v}}^{(i)}))/\tau\right)$, and $\text{sim}(\mathbf{u}, \mathbf{w}) = \mathbf{u}^\top \mathbf{w}$ denotes the cosine similarity between the ℓ_2-normalized vectors $\mathbf{u}$ and $\mathbf{w}$, τ is a temperature parameter, and $\bar{v}$ defines a view $\neq v$. The pair of normalized embeddings $(\mathbf{z}_f^{(i)}, \mathbf{z}_l^{(i)})$ from the same study i is treated as a positive pair where $\mathbf{z}_v^{(i)} = f_\theta(\mathbf{x}_v^{(i)})$. Embeddings from different studies $j \neq i$ regardless of the view (i.e., $\mathbf{z}_f^{(j)}$ and $\mathbf{z}_l^{(j)}$) serve as negative samples. In distributed training, negatives are gathered across devices for a more expressive representation without extra annotation burden.

4. Experiments

We aimed to investigate how the representations learned through our MVMAE models and the contrastive learning approach influence performance on a downstream classification task under varying levels of label availability. This setting mimics realistic clinical practice, where expert annotation is scarce and expensive. We study two complementary questions: (i) *How effective are the representations learned by a pretrained encoder when the model is fully fine-tuned on a downstream classification task?* In particular, we assess whether leveraging the data structure during pretraining leads to measurable improvements. (ii) *To what extent is fine-tuning necessary? Can we reach comparable performance via linear probing?* For this, we would keep the encoder frozen and train only a linear classifier.

We compare three training regimes that share the same ViT-b backbone (Dosovitskiy et al., 2020) and optimizer but differ in how that backbone is initialized: the proposed MV-MAE; the proposed contrastive-only variant that drops the reconstruction loss; a purely supervised baseline trained from scratch. All self-supervised models are first exposed to the full unlabeled training pool; downstream experiments are then performed on splits of 5K, 10K, 25K, 50K, and the complete 102K labeled studies drawn from the training set. We then compute the numbers over the full validation set.

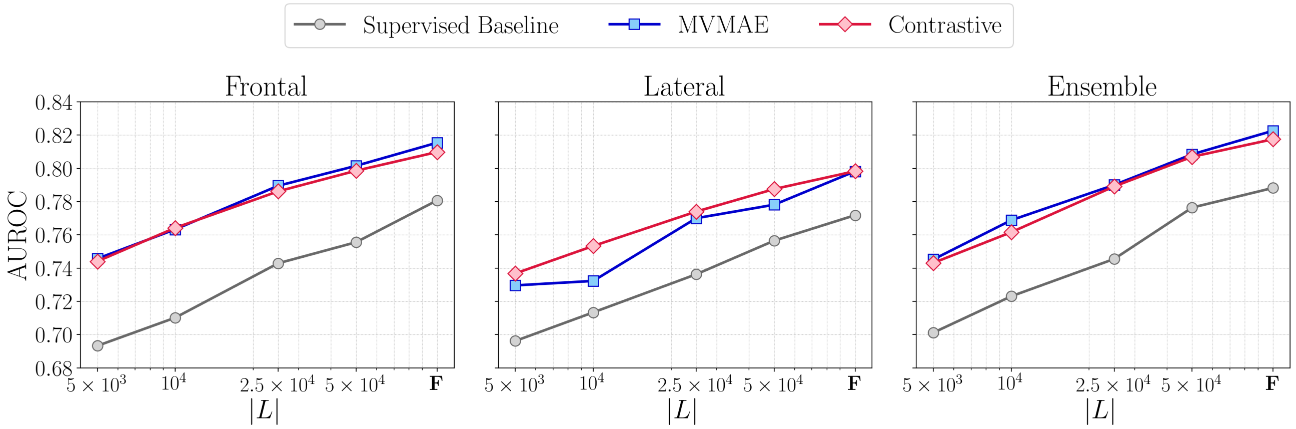

Figure 3. Performance comparison of MVMAE, Contrastive, and Supervised methods across three evaluation settings: (a) *Frontal*, (b) *Lateral*, and (c) *Ensemble*. Each plot shows the AUROC score, computed as a macro-average across 14 pathology labels, achieved by each method under varying numbers of labeled samples $|L|$. **F** denotes the total dataset size (10.2×10^4).

All models were trained using data augmentations common in this type of learning paradigm: random resize cropping, horizontal clipping, and jitter addition to improve the prediction and representation capabilities. We perform this analysis through two experiments.

Experiment 1 assesses *label-efficiency under end-to-end fine-tuning*. After attaching a randomly initialized linear classification head, we unfreeze all the encoder weights and train them on each supervision tier, keeping learning-rate schedules identical so that any performance difference can be attributed to the representation rather than to additional compute. Monitoring the AUROC (macro-averaged over all labels) as the number of labels grows allows us to quantify how quickly each pretraining strategy narrows the gap to its fully supervised counterpart. Results are reported in Figure 3 evaluated across three modality scenarios: the set of *Frontal* views, *Lateral* views, and an *Ensemble* of both.

Experiment 2 probes *representation quality with the encoder frozen*. Here, we restrict supervision to the smallest 5K-label subset and compare the full fine-tuning approach from Experiment 1 to a linear probing approach with only a trainable linear classifier head. Because encoder weights remain fixed in the latter, any improvement must originate from the quality of the features rather than from further optimization of the backbone. Results are summarized in Table 1 right panel.

Note that the ensemble is built as a late fusion of the unimodal single view scores. All models are selected on a held-out validation set and reported on the test set using macro-AUROC over the fourteen CheXpert labels.

Discussion From our results, we see that (1) Multimodal pretraining consistently improves downstream classification performance compared to training from scratch. (2) Leveraging data structure during pretraining proves more effective than enforcing structure at the supervision stage only (i.e., via an ensemble supervised classifier). Notably, unimodal performance of the pretrained models exceeds or matches that of the supervised ensemble classifier, suggesting that soft information sharing during pretraining is an effective way to exploit the underlying data structure. (3) Fine-tuning the encoders previously learnt proves more useful than directly probing the representations.

Strategy	Modality	Fine-tuning	Linear Probing
Supervised	Frontal	0.69	-
	Lateral	0.70	-
	Ensemble	0.70	-
MVMAE	Frontal	0.75	0.65
	Lateral	0.73	0.65
	Ensemble	0.75	0.69
Contrastive	Frontal	0.74	0.65
	Lateral	0.74	0.64
	Ensemble	0.74	0.66

Table 1. **Comparison of fine-tuning the pretrained encoder vs linear probing across methods and modality types on the classification task.** Each method (Supervised, MVMAE, Contrastive) is evaluated using different types (Frontal, Lateral, Ensemble) and trained only on 5000 labeled samples.

5. Conclusion

We present a multi-view regularized masked autoencoder that leverages the natural structure of clinical imaging data—specifically, paired anatomical views—to learn robust, informative representations without requiring textual supervision. By combining masked reconstruction with cross-view alignment, our approach enables efficient pretraining leveraging the inherent structure of medical datasets. Experiments on MIMIC-CXR demonstrate that this strategy closes much of the gap to vision–language models while remaining lightweight and domain-adaptable. Looking forward, the framework is broadly applicable to other structured medical modalities, such as longitudinal studies or multi-sequence scans, offering a scalable path toward foundation models grounded in the realities of clinical data.

Acknowledgments

This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a02 on Alps, and by the LOOP Zurich as part of the application driver project supporting the LOOP Zurich Biomedical Informatics Platform (BMIP). MV and SL are supported by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00047. AR is supported by the StimuLoop grant #1-007811-002 and the Vontobel Foundation. ND and FN received research support from the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH)-Rapid Action Call - under TRUST-RAD project. MK is supported by the UZH Global Strategy and Partnerships Funding Scheme and a Research Partnership Grant with China, Japan, South Korea and the ASEAN region (RPG 072023 18). TS is supported by the grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology).

References

Agostini, A., Chopard, D., Meng, Y., Fortin, N., Shahbaba, B., Mandt, S., Sutter, T. M., and Vogt, J. E. Weakly-supervised multimodal learning on mimic-cxr. *arXiv preprint arXiv:2411.10356*, 2024.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

Chaves, J. M. Z., Huang, S.-C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Chen, Z., Varma, M., Delbrouck, J., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J., Youssef, A., Cohen, J. P., Reis, E., et al. Chexagent: Towards a foundation model for chest x-ray interpretation, arxiv, 2024. *arXiv preprint arXiv:2401.12208*, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Haque, M. I. U., Dubey, A. K., Danciu, I., Justice, A. C., Ovchinnikova, O. S., and Hinkle, J. D. Effect of image resolution on automated classification of chest X-rays. *Journal of Medical Imaging*, 10(4):044503–044503, 2023.

Hassanzadeh, H., Kholghi, M., Nguyen, A., and Chu, K. Clinical document classification using labeled and unlabeled data across hospitals. In *AMIA annual symposium proceedings*, volume 2018, pp. 545, 2018.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. MIMIC-CXR-JPG-chest radiographs with structured labels. *PhysioNet*, 2019a.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019b.

Mo, S. and Liang, P. P. Multimed: Massively multimodal and multitask medical understanding. *arXiv preprint arXiv:2408.12682*, 2024.

Pellegrini, C., Özsoy, E., Busam, B., Wiestler, B., Navab, N., and Keicher, M. Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In *Medical Imaging with Deep Learning*, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., and Rosenow III, E. C. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

Sutter, T., Meng, Y., Agostini, A., Chopard, D., Fortin, N., Vogt, J., Shahbaba, B., and Mandt, S. Unity by diversity: Improved representation learning for multimodal vaes. *Advances in Neural Information Processing Systems*, 37: 74262–74297, 2024.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and Rajpurkar, P. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Xiao, J., Bai, Y., Yuille, A., and Zhou, Z. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600, 2023.