# Bidirectional Anatomy-Aware Post-Training for Longitudinal Chest X-Ray Progression Modeling

Yuming Chen[1], Hui Wang[1], Minh-Son To[3], Jianpeng Zhang[4], Qi Wu[1], and Yutong Xie[2✉]

[1] Adelaide University
[2] Mohamed bin Zayed University of Artificial Intelligence
[3] Flinders University
[4] Zhejiang University
yuming.chen01@adelaide.edu.au, yutong.xie@mbzuai.ac.ae

**Abstract.** Longitudinal comparison is a routine yet clinically consequential component of chest X-ray (CXR) interpretation, informing disease progression assessment and treatment decisions. Prior temporal CXR models are typically trained as general-purpose pair encoders and adapted with standard progression classifiers, but they do not explicitly capture two key properties of clinical progression assessment: temporal symmetry (for reversible findings, the progression assessment should invert when the temporal order is reversed) and anatomical grounding (decisions rely on localized evidence within specific structures). We propose Bidirectional Anatomy-Aware Progression Perception (BAAP), a lightweight post-training strategy built on top of a pretrained CXR encoder. BAAP encodes temporal symmetry via bidirectional pair augmentation, jointly optimizing each CXR pair in both chronological orders with label inversion to enforce directional consistency and mitigate directional bias. To ground predictions anatomically, we introduce an Anatomy-aware Perception Layer (APL) that aggregates patch-level features within predefined anatomical regions and fuses prior and current region features with their differences for per-region progression classification. Adding only 0.63M parameters (≈0.7% of a ViT-B/16 backbone) and requiring no additional annotations, BAAP achieves 65.90% mean accuracy on MS-CXR-T (+4.78% over the strongest baseline) and 66.4% mean anatomy-level accuracy on Chest ImaGenome Gold (+11.0% over the strongest baseline), demonstrating that a clinically motivated symmetry prior and anatomy-aware design can yield substantial gains for longitudinal CXR progression prediction with minimal overhead. The code is available at `https://github.com/chenyuming052/BAAP`.

**Keywords:** Longitudinal Chest X-ray · Progression Prediction · Post-training.

## 1 Introduction

When prior examinations are available, longitudinal comparison is a routine and clinically consequential part of CXR interpretation [16]. Radiologists evaluate

how findings evolve over time, for example, whether a pleural effusion is resolving, consolidation is progressing, or edema remains stable. These temporal judgments directly guide treatment adjustment and discharge planning, making progression assessment an essential component of the thoracic radiology workflow [10]. Two properties of this assessment are worth highlighting. First, progression judgment is *temporally symmetric*: for the reversible findings that dominate acute thoracic pathology, a finding that improves between two time points must correspondingly worsen when the temporal order is reversed. Second, it is *anatomically grounded*: radiologists inspect specific structures (e.g., costophrenic angles, hila, and cardiac silhouette) and reason about localized changes rather than forming a single global impression of the image [3].

Recent self-supervised learning and vision–language pretraining have substantially improved CXR representation learning [25,8,19,2,4]. Temporal extensions [1,23,5,24,12,13] further incorporate longitudinal information by learning from image pairs acquired at different time points. However, such models are typically adapted with standard progression-classification formulations, leaving the two properties above only partially addressed: even recent order-inversion methods [11,6] operate at the whole-image level and lack explicit per-anatomy localized progression classification. Specifically, two key gaps remain: (1) Temporal symmetry. Common training setups assume a fixed chronological (prior→current) ordering without enforcing consistency under temporal reversal, leaving symmetry cues underutilized and potentially introducing directional bias. (2) Anatomical grounding. Progression is commonly inferred from holistic image-pair representations, while explicitly localizing and aggregating evidence within anatomical structures remains less explored, which may dilute subtle yet clinically important region-specific changes.

To address these limitations, we propose Bidirectional Anatomy-Aware Progression Perception (**BAAP**), a lightweight post-training strategy built on top of a pretrained image encoder. BAAP consists of two complementary components. First, a Bidirectional Progression Training paradigm that augments each training pair with its temporally reversed counterpart and the correspondingly inverted label. Joint optimization on both orderings explicitly encodes temporal symmetry as a structural prior, reducing directional bias and improving robustness without requiring additional annotations. Second, an Anatomy-aware Perception Layer (APL) that shifts progression prediction from global image-level embeddings to anatomically localized evidence. APL pools patch-level features within predefined anatomical regions, computes region-level temporal difference features, and fuses these signals for per-anatomy progression classification. This design better reflects radiologists' longitudinal reading behavior and improves sensitivity to localized changes.

By combining bidirectional training with anatomy-aware perception, BAAP enables temporally symmetric and anatomically grounded progression modeling. Notably, BAAP is lightweight, adding only 0.63 M parameters (≈0.7% of the ViT-B backbone), demonstrating that substantial gains can be achieved through task-aligned design rather than increased model capacity. Extensive experiments

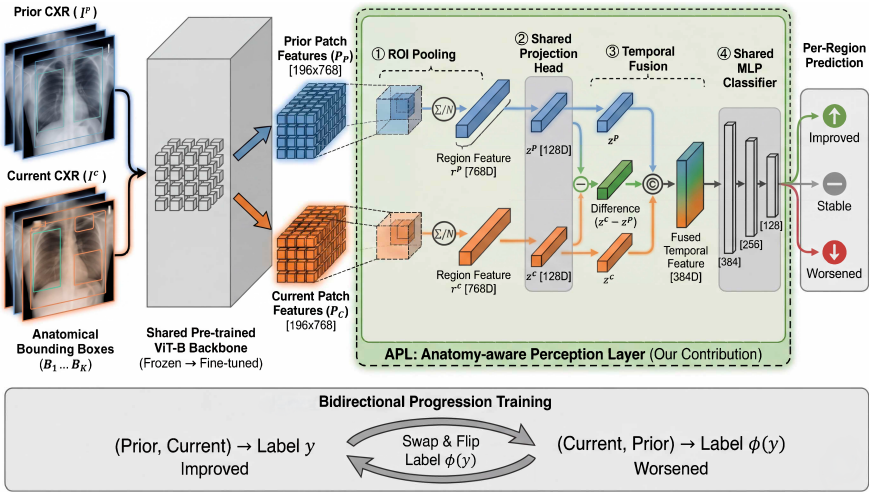


**Fig. 1.** Overview of the BAAP post-training strategy. Given a pretrained CXR image encoder, the lightweight APL pools patch features within anatomy bounding boxes, projects them to a compact embedding, computes temporal difference features, and classifies per-region progression. Bidirectional training reverses temporal order with label inversion $\phi$, augmenting each pair with its reversed ordering and promoting directional consistency.

on the MS-CXR-T benchmark [1] and the Chest ImaGenome Gold dataset [22] demonstrate the effectiveness of BAAP. On MS-CXR-T, BAAP achieves 65.90% average accuracy (+4.78% over the baseline) and on Chest ImaGenome Gold, 66.4% mean anatomy-level accuracy (+11.0% over the strongest baseline). Our contributions are summarized as follows:

- We propose a **Bidirectional Progression Training** paradigm that encodes temporal symmetry as a structural prior, reducing directional bias in longitudinal CXR modeling.
- We introduce an **Anatomy-aware Perception Layer (APL)** that captures fine-grained, region-level progression cues through anatomy-guided patch aggregation and temporal difference modeling.
- We develop a **lightweight post-training strategy** that adds around 0.7% more parameters to the backbone while establishing a new state of the art on MS-CXR-T and Chest ImaGenome Gold benchmarks.

## 2 Method

Given a pretrained CXR image encoder (we adopt the competitive self-supervised MedST [23]), BAAP comprises two complementary components: *Bidirectional Progression Training* (section 2.2) and an *Anatomy-Aware Perception Layer* (APL, section 2.3), capturing fine-grained progression cues while preserving anatomical structure. An overview of BAAP is shown in Fig. 1.

### 2.1 Problem Formulation

Given a temporal CXR pair with a prior image $I^p$ and a current image $I^c$, together with per-anatomy boxes $\{(b_k^p, b_k^c)\}_{k=1}^K$ (from a predefined set of up to 38 anatomical structures [22]), the goal is to predict an $X$-way progression label $y_k \in \{\textit{improved}, \textit{stable}, \textit{worsened}\}$ for each annotated region $k$, where $X = 3$. We learn a classifier $f$ that outputs logits over $X$ classes, and predict $\hat{y}_k = \arg\max f(I^p, I^c, b_k^p, b_k^c)$. Let $\mathcal{D} = \{(I_i^p, I_i^c, \{(b_{i,k}^p, b_{i,k}^c, y_{i,k})\}_{k=1}^{K_i})\}_{i=1}^{|\mathcal{D}|}$ denote the training set.

### 2.2 Bidirectional Progression Training

Clinical progression assessment is often approximately temporally symmetric for reversible findings: if a finding improves from time $t_1$ to $t_2$, it typically corresponds to worsening when the order is reversed. We encode this invariance as a structural prior via bidirectional pair augmentation, swapping temporal order and flipping the progression labels. For each original sample $(I^p, I^c, y_k)$, we construct a temporally reversed sample $(I^c, I^p, \phi(y_k))$ by swapping the prior and current images and applying the label inversion function:

$$\phi(y) = \begin{cases} \textit{worsened}, & \text{if } y = \textit{improved}, \\ \textit{improved}, & \text{if } y = \textit{worsened}, \\ \textit{stable}, & \text{if } y = \textit{stable}. \end{cases} \tag{1}$$

Jointly training on both temporal orderings is achieved by optimizing the bidirectional cross-entropy loss:

$$\mathcal{L}_{bidir} = CE(f(I^p, I^c, b_k^p, b_k^c), y_k) + CE(f(I^c, I^p, b_k^c, b_k^p), \phi(y_k)) \tag{2}$$

where CE$(\cdot)$ denotes cross-entropy.

Since each *improved* instance yields a paired *worsened* instance (and vice versa), the augmentation also alleviates imbalance between the two directional classes. In practice, this reduces directional bias and improves robustness without additional annotations.

### 2.3 Anatomy-Aware Perception Layer

Both images are encoded by a shared ViT-Base backbone [7]. Given an input of size $H \times W$, the backbone partitions it into a $G \times G$ grid of non-overlapping $S \times S$ patches and produces token embeddings. We discard the `[CLS]` token and keep the $N = G^2$ patch tokens as grid-aligned features:

$$\mathbf{P}_p = \text{ViT}(I^p), \qquad \mathbf{P}_c = \text{ViT}(I^c), \tag{3}$$

where $\mathbf{P}_p, \mathbf{P}_c \in \mathbb{R}^{N \times D}$. In our setting, $H = W = 224$, $P = 16$, $G = 14$, hence $N = 196$ and $D = 768$. APL predicts progression from anatomically localized evidence instead of a single global descriptor, using four stages described below.

**ROI Pooling.** For each region $k$, we use the anatomy boxes $b_k^p$ and $b_k^c$ to select the patch tokens inside the region for the prior and current images, respectively. This selection is represented by binary masks $\mathbf{M}_k^p, \mathbf{M}_k^c \in \{0,1\}^N$, where $M_{k,n}^{(\cdot)} = 1$ if token $n$ falls inside $b_k^{(\cdot)}$ (and 0 otherwise). We obtain region features by masked average pooling:

$$\mathbf{r}_k^{(\cdot)} = \frac{\sum_{n=1}^N M_{k,n}^{(\cdot)} \, \mathbf{p}_n^{(\cdot)}}{\sum_{n=1}^N M_{k,n}^{(\cdot)}} \in \mathbb{R}^D, \qquad (\cdot) \in \{p, c\}, \tag{4}$$

where $\mathbf{p}_n^{(\cdot)} \in \mathbb{R}^D$ denotes the $n$-th patch token from $\mathbf{P}^{(\cdot)}$.

**ROI Projection.** The pooled region-of-interest (ROI) feature $\mathbf{r}_k^{(\cdot)} \in \mathbb{R}^D$ is mapped to a compact embedding $\mathbf{z}_k^{(\cdot)} \in \mathbb{R}^d$ using a shared two-layer MLP. Specifically, we apply a linear layer ($D \to h$), followed by BatchNorm and GELU, and then a second linear layer ($h \to d$) with BatchNorm. Finally, we $\ell_2$-normalize the output embedding before temporal fusion. The projection head is shared across regions and between the prior and current branches.

**Temporal difference fusion.** Given the projected ROI embeddings $\mathbf{z}_k^p, \mathbf{z}_k^c \in \mathbb{R}^d$, we form a progression representation by concatenating the two states and their directional difference:

$$\mathbf{u}_k = \left[ \, \mathbf{z}_k^p \parallel \mathbf{z}_k^c \parallel (\mathbf{z}_k^c - \mathbf{z}_k^p) \, \right] \in \mathbb{R}^{3d}. \tag{5}$$

where $[\cdot \| \cdot]$ denotes concatenation along the channel dimension. The difference term encodes directional temporal variation, while retaining both $\mathbf{z}_k^p$ and $\mathbf{z}_k^c$ preserves absolute appearance information from each time point.

**Shared MLP classifier.** We use a single MLP classifier shared across all regions to predict the progression label from the fused representation:

$$\hat{y}_k = \arg\max_X \ \mathrm{MLP}(\mathbf{u}_k)[X], \qquad \mathrm{MLP} : 3d \to h_1 \to h_2 \to X, \tag{6}$$

where the MLP parameters are shared across regions.

## 3 Experiments

### 3.1 Datasets and Evaluation

**MIMIC-CXR.** We initialize our backbone on MIMIC-CXR [9], a large-scale public dataset comprising 377,110 chest radiographs from 65,379 patients, each paired with a free-text radiology report. Pre-training follows the MedST [23] recipe.

**Chest ImaGenome.** For anatomy-level progression supervision, we use Chest ImaGenome [22], a MIMIC-CXR–derived resource that augments each frontal radiograph with anatomy-centric scene graphs, including bounding boxes for 38 predefined regions and region-wise temporal comparison labels (*improved*, *worsened*, *no_change*) between sequential exams. We post-train on the Silver split

**Table 1.** Disease-level progression accuracy (%) on the MS-CXR-T benchmark for five findings. Results are averaged over three runs with different random seeds and reported as mean ± standard deviation.

| Model | Consol. | Edema | Pl. Eff. | Pneum. | Pneumoth. | Avg. |
|---|---|---|---|---|---|---|
| *Single-image methods* | | | | | | |
| ConVIRT [25] | 46.6±1.0 | 57.0±1.0 | 54.5±0.2 | 61.1±1.5 | 48.5±0.2 | 53.6 |
| GLoRIA-C [8] | 49.1±1.5 | 53.7±0.6 | 49.6±1.8 | 58.1±2.2 | 45.6±2.5 | 51.2 |
| GLoRIA-M [8] | 45.0±0.5 | 49.1±1.5 | 47.9±3.1 | 60.2±1.1 | 41.5±0.9 | 48.7 |
| MGCA [19] | 50.8±0.4 | 62.7±0.2 | 57.2±0.7 | 63.7±0.9 | 52.2±1.0 | 57.3 |
| MedCLIP [21] | 50.3±0.7 | 54.5±2.8 | 56.6±0.7 | 58.9±2.6 | 45.2±1.5 | 53.1 |
| MRM [20] | 46.3±2.9 | 49.1±5.7 | 49.1±0.7 | 55.1±1.4 | 45.5±3.0 | 49.0 |
| BioViL [2] | 56.4±0.2 | 57.3±0.8 | 54.5±0.4 | 67.1±0.0 | 55.3±0.2 | 58.1 |
| *Temporal methods* | | | | | | |
| BioViL-T [1] | 56.9±1.8 | 61.6±1.0 | 53.9±0.9 | 67.2±0.2 | 55.5±0.0 | 59.0 |
| Med-ST [23] | 60.6±1.2 | 67.4±0.3 | 58.5±1.5 | 65.0±0.3 | 54.2±0.8 | 61.1 |
| **BAAP (Ours)** | **64.3**±1.2 | **70.9**±0.4 | **68.7**±1.5 | **70.5**±0.6 | **55.5**±0.2 | **65.9** |

and prevent any overlap with MS-CXR-T [1] by removing all MS-CXR-T study pairs from the Silver training pool. After quality filtering and constructing bidirectional temporal pairs, we obtain approximately 252,000 region-level training samples. We evaluate on the Gold split, where expert radiologists verified annotations for 500 patients across 23 of the 38 regions.

**MS-CXR-T.** We evaluate disease-level temporal progression on MS-CXR-T [1], which contains 1,326 radiologist-annotated longitudinal CXR pairs from 800 subjects spanning five thoracic findings (consolidation, edema, pleural effusion, pneumonia, pneumothorax) with three-way labels (*improved*, *stable*, *worsened*). MS-CXR-T is curated by selecting candidate pairs from the Chest ImaGenome Silver split after excluding studies previously verified in Chest ImaGenome Gold, and further discarding pairs with conflicting progression labels for the same finding to retain unambiguous supervision.

**Evaluation protocol.** We evaluate disease-level progression on MS-CXR-T and report mean±std accuracy. On Chest ImaGenome Gold, we report per-region anatomy-level accuracy in Table 2. For concise reporting, the 23 annotated regions are grouped into five clinically meaningful categories based on anatomical proximity (e.g., Lung Sub-regions, Hilum, Mediastinal & Cardiac), with per-group accuracy averaged over constituent regions.

**Implementation details.** We adopt a ViT-Base backbone initialized from DeiT-pretrained weights [18] and pretrained using the MedST framework [23]. The model takes $224 \times 224$ inputs on a $14 \times 14$ patch grid. The ROI projection maps 768-dim patch features to 128 dimensions; the fusion layer concatenates prior, current, and difference embeddings (384-dim total) and feeds into a 3-layer MLP classifier (384→256→128→3). We train for 15 epochs with AdamW [15]

**Table 2.** Anatomy-level temporal progression classification accuracy (%) on the Chest ImaGenome Gold dataset. Med./Card. = Mediastinal & Cardiac; Pleu./Dia. = Pleural & Diaphragm.

| Method | Lung Sub-Regions | Whole Lung | Hilum | Med./Card. | Pleu./Dia. | Avg. |
|---|---|---|---|---|---|---|
| *Single-image methods* | | | | | | |
| BioViL [2] | 48.0 | 47.1 | 44.9 | 72.4 | 54.7 | 53.4 |
| MGCA [19] | 50.9 | 48.6 | 52.7 | 76.6 | 43.8 | 54.5 |
| *Temporal methods* | | | | | | |
| MedGemma1.5 [17] | 52.3 | 53.9 | 43.1 | 78.0 | 41.6 | 53.8 |
| BioViL-T [1] | 47.3 | 48.0 | 50.9 | 77.6 | 53.3 | 55.4 |
| MedST [23] | 54.5 | 48.0 | 47.3 | 77.1 | 48.9 | 55.1 |
| **BAAP (Ours)** | **56.9** | **61.3** | **68.9** | **86.0** | **59.1** | **66.4** |

(lr=$10^{-3}$, weight decay $10^{-4}$), cosine annealing [14] with 10% warmup, batch size 10 per GPU, class-weighted cross-entropy loss with label smoothing ($\epsilon$=0.1). The backbone is frozen for the first 5 epochs, then unfrozen with a 0.01× learning rate multiplier for staged fine-tuning. All experiments are conducted on two NVIDIA A100 (80 GB) GPUs using PyTorch Distributed Data Parallel (DDP).

### 3.2 Results

**Disease-Level Progression Classification.** We first evaluate disease-level progression prediction on MS-CXR-T [1], where each longitudinal CXR pair is annotated with disease progression labels for the findings present in that pair. As shown in Table 1, BAAP achieves an overall accuracy of 65.9%, outperforming the MedST baseline (61.1%) by +**4.8%**. Notably, these performance gains are particularly pronounced for conditions that heavily rely on local spatial context, such as Pleural Effusion (+10.2%) and Pneumonia (+5.5%). The improvement is consistent across evaluated findings, indicating that anatomy-aware post-training on localized regions yields temporal representations that translate effectively to robust disease-level progression prediction. Furthermore, by explicitly grounding temporal differences within specific anatomical regions, BAAP successfully mitigates the confounding effects of patient positioning and breathing variations across longitudinal scans, enabling the model to focus on genuine pathological changes.

**Anatomy-Level Progression Classification.** Table 2 reports the anatomy-level temporal progression classification accuracy on the Chest ImaGenome Gold Dataset. Specifically, BAAP achieves an average accuracy of 66.4%, surpassing the strongest baseline BioViL-T (55.4%) by 11.0 %. The improvements are particularly pronounced in the Hilum region (+16.2% over MGCA) and the Whole Lung region (+7.4% over MedGemma1.5), where prior methods show limited discriminative ability. For the Mediastinal & Cardiac region, where all baselines

**Table 3.** Cumulative ablation study under the BAAP strategy. Uni-PT: Unidirectional Progression Post-Training; Bidir.: Bidirectional Progression Perception Post-Training; APL: Anatomy-aware Perception Layer. MS-CXR-T reports per-disease and average accuracy (%).

| | Components | | | MS-CXR-T | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Uni-PT | Bidir. | APL | Consol. | Edema | Pl.Eff. | Pneum. | Pneurx. | Avg. | $\Delta$ |
| Baseline | – | – | – | 60.6 | 67.4 | 58.5 | 65.0 | 54.2 | 61.1 | – |
| + Uni-PT | ✓ | – | – | 61.2 | 67.8 | 68.4 | 67.1 | 55.3 | 64.0 | +2.9 |
| + Bidir. | ✓ | ✓ | – | 63.8 | 72.2 | 68.9 | 65.6 | 55.5 | 65.2 | +1.2 |
| BAAP (Ours) | ✓ | ✓ | ✓ | 64.3 | 70.9 | 68.7 | 70.5 | 55.5 | **65.9** | +0.7 |

already achieve relatively high accuracy due to the more salient visual changes, BAAP still improves upon the best competitor MedGemma by 8.0%. These results demonstrate that BAAP can effectively capture fine-grained temporal representations at the anatomical level, enabling accurate progression classification even without supervision on the target evaluation set.

### 3.3 Ablation Study

We conduct a cumulative ablation under the BAAP strategy (Table 3). Adding Uni-PT to MedST improves disease-level accuracy by +2.9% on MS-CXR-T, indicating that explicit temporal-ordering supervision improves direction-sensitive prediction. Incorporating bidirectional training yields further gains of +1.2% on MS-CXR-T, suggesting that enforcing temporal symmetry captures additional progression cues. Adding APL achieves the best performance, reaching 65.9% on MS-CXR-T.

## 4 Conclusion

We proposed BAAP, a lightweight post-training strategy that encodes temporal symmetry via bidirectional pair augmentation and anatomical grounding via the Anatomy-Aware Perception Layer. With approximately 0.7% additional parameters, BAAP improves the MedST baseline on MS-CXR-T by 4.78% to reach 65.90% average accuracy, and improves BioViL-T on Chest ImaGenome Gold by 11.0% to reach 66.4% anatomy-level accuracy. These consistent gains across both disease-level and anatomy-level evaluations suggest that task-aligned inductive biases can be more effective than scaling model capacity for longitudinal CXR progression prediction. A current limitation is the reliance on predefined anatomical bounding boxes, though off-the-shelf anatomy detectors can supply the required per-region boxes when such annotations are unavailable. Future work will incorporate automated anatomy localization end-to-end and extend BAAP from pairwise comparison to multi-visit longitudinal sequences.

**Disclosure of Interests.** The authors have no competing interests to declare.

## References


1. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15016–15027 (June 2023)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
3. Chen, Q., Xie, Y., Wu, B., Chen, X., Ang, J., To, M.S., Chang, X., Wu, Q.: Act like a radiologist: radiology report generation across anatomical regions. In: Proceedings of the Asian Conference on Computer Vision. pp. 1–17 (2024)
4. Chen, Q., Zhao, R., Wang, S., Phan, V.M.H., Hengel, A.v.d., Verjans, J., Liao, Z., To, M.S., Xia, Y., Chen, J., et al.: A survey of medical vision-and-language applications and their techniques. arXiv preprint arXiv:2411.12195 (2024)
5. Chen, Y., Xu, S., Sellergren, A., Matias, Y., Hassidim, A., Shetty, S., Golden, D., Yuille, A.L., Yang, L.: CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15965. Springer Nature Switzerland (September 2025)
6. Choi, D., Ko, H., Lee, J.H., Park, C.M.: A bidirectional loss approach to imparting order sensitivity to multi-image chest x-ray encoders. In: Medical Imaging with Deep Learning - Short Papers (2025)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
8. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
10. Karwande, G., Mbakwe, A.B., Wu, J.T., Celi, L.A., Moradi, M., Lourentzou, I.: Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 581–591. Springer (2022)
11. Ko, H., Jeon, K., Choi, D., Park, C.M.: Temporal inversion for learning interval change in chest x-rays. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28338–28347 (2026)
12. Lian, C., Zhou, H.Y., Liang, D., Qin, J., Wang, L.: Efficient medical vision-language alignment through adapting masked vision models. IEEE Transactions on Medical Imaging (2025)
13. Liu, C., Yao, W., Yin, K., Cheung, W.K., Qin, J.: Multimodal disease progression modeling via spatiotemporal disentanglement and multiscale alignment. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)

14. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (2017)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Loy, C.T., Irwig, L.: Accuracy of Diagnostic Tests Read With and Without Clinical InformationA Systematic Review. JAMA **292**(13), 1602–1609 (Oct 2004). https://doi.org/10.1001/jama.292.13.1602
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T., Sterling, B., Baby, S.A., Baby, S.M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Riviere, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J.B., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J., Warkentin, T., Shlens, J., Fleet, D., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D.F., Kirmizibayrak, C., Pilgrim, R., Golden, D., Yang, L.: Medgemma technical report (2025)
18. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
19. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in neural information processing systems **35**, 33536–33549 (2022)
20. Wang, Y., Zhao, Y., Petzold, L.: Are large language models ready for healthcare? a comparative study on clinical language understanding. In: Deshpande, K., Fiterau, M., Joshi, S., Lipton, Z., Ranganath, R., Urteaga, I., Yeung, S. (eds.) Proceedings of the 8th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 219, pp. 804–823. PMLR (11–12 Aug 2023)
21. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
22. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316 (2021)
23. Yang, J., Su, B., Zhao, X., Wen, J.R.: Unlocking the power of spatial and temporal information in medical multimodal pre-training. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 56382–56396. PMLR (21–27 Jul 2024)
24. Yang, Z., Shen, L.: Tempa-vlp: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4625–4634. IEEE (2025)
25. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25. PMLR (2022)