CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding

Yixiong Chen[1⋆], Shawn Xu[2], Andrew Sellergren[2], Yossi Matias[2], Avinatan Hassidim[2], Shravya Shetty[2], Daniel Golden[2], Alan L. Yuille[1], and Lin Yang[2]

[1]Johns Hopkins University, [2]Google Research

Abstract. Vision-language models have proven to be of great benefit for medical image analysis since they learn rich semantics from both images and reports. Prior efforts have focused on better alignment of image and text representations to enhance image understanding. However, though explicit reference to a prior image is common in Chest X-Ray (CXR) reports, aligning progression descriptions with the semantics differences in image pairs remains under-explored. In this work, we propose two components to address this issue. (1) A CXR report processing pipeline to extract temporal structure. It processes reports with a large language model (LLM) to separate the description and comparison contexts, and extracts fine-grained annotations from reports. (2) A contrastive captioner model for CXR, namely CoCa-CXR, to learn how to both describe images and their temporal progressions. CoCa-CXR incorporates a novel regional cross-attention module to identify local differences between paired CXR images. Extensive experiments show the superiority of CoCa-CXR on both progression analysis and report generation compared to previous methods. Notably, on MS-CXR-T progression classification, CoCa-CXR obtains 65.0% average testing accuracy on five pulmonary conditions, outperforming the previous state-of-the-art (SOTA) model BioViL-T by 4.8%. It also achieves a RadGraph F1 of 24.2% on MIMIC-CXR, which is comparable to the Med-Gemini foundation model.

Keywords: Vision Language Models · Progression Prediction · Report Generation.

1 Introduction

Recent advances in vision-language (VL) pre-training have significantly enhanced the development of flexible and powerful models for chest X-ray (CXR) analysis. Training on both images and reports allows models to capture rich semantics, aligning medical concepts with image representations. However, CXR reports frequently include comparisons between multiple examinations [13], a crucial

⋆ This work was done during an internship at Google. Corresponding author: ychen646@jh.edu

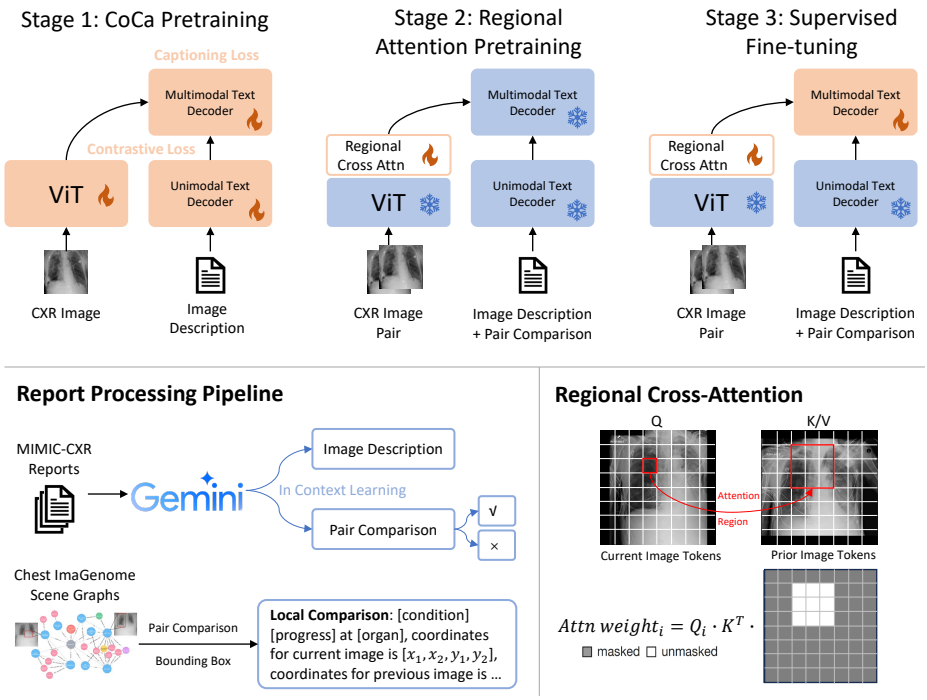

Fig. 1: CoCa-CXR adds a regional cross-attention module to CoCa and is trained with three stages. We utilize an LLM, Gemini, to parse MIMIC-CXR reports and get image description and pair comparison. Chest ImaGenome scene graphs provide us with the local comparison between CXR image pairs. To leverage the locality of disease conditions, we apply cross-attention to emphasize the correlation between neighboring tokens of two images.

aspect often overlooked by existing approaches [4, 11, 19, 21]. Most prior work focuses on predicting findings from a single image without explicitly modeling temporal progression, limiting their ability to understand disease evolution and restricting their clinical applicability.

Recent literature [1, 2, 7, 23, 26] has begun to train VL models that can make temporal predictions with multiple CXR images. The common practice is fusing representations of images as the joint representation to generate reports. While straightforward, it does not explicitly guide models to learn temporal differences. The key challenges remain: (1) the lack of datasets with aligned image pairs and comparative descriptions, and (2) the need for model architectures to effectively capture subtle regional changes over time.

We propose a systematic framework (Fig. 1) to address the above problems. For (1), we build a CXR report processing pipeline to curate a new dataset, CXR-4, with four sub-datasets, detailed in Tab. 1. It uses a large language model (LLM) [20] to separate reports into descriptions and comparisons, allowing the model to learn them sequentially. In addition, we also leverage the comparison

Table 1: Statistics of the sub-datasets used for training CoCa-CXR.

Sub-Dataset	Stage	Description	Data Source	#Samples
1. Clean image-report pair	1&2&3	Image-report pairs without any image pair comparison content.	MIMIC-CXR	224,487
2. Image pair & filtered report	2&3	Image pair with corresponding report. All reports must contain comparison info.	MIMIC-CXR	132,320
3. Image pair & comparison info	2&3	Image pair with corresponding comparison sentences from the report.	MIMIC-CXR	259,562
4. Image pair & abnormal organs	2&3	Image pair with abnormal organ's condition, coordinates, and progression.	MIMIC-CXR & Chest ImaGenome	758,344

and localization from scene graphs [25] which contain the description of the abnormal organs, condition progressions, and the corresponding bounding boxes, enabling the models to learn regional differences. For (2), to leverage the regional comparison in the CXR-4 dataset, we propose regional cross-attention, inspired by [6], but specifically designed for attention between current and prior images. This module refines traditional cross-attention [22] by restricting each token in an image to attend only to its surrounding tokens in the prior image. We choose Contrastive Captioner (CoCa) [28] as our baseline model. It has a minimalist architecture to perform VL contrastive and generative learning together, which are essential for aligned representation and downstream multitasking.

In summary, our work presents three major contributions to temporal CXR understanding. 1) We introduce the CXR-4 dataset. It provides not only the alignment between CXR images and text descriptions, but also explicit comparison. 2) We propose CoCa-CXR, which is a CoCa-based model that can generate reports from image pairs, predicting condition progressions, and localizing abnormal organs. 3) Experiments on both progression prediction and report generation tasks show the superior performance of CoCa-CXR to previous SOTA CXR temporal models.

2 CXR-4 Dataset

We introduce CXR-4, a new CXR dataset comprising four sub-datasets (Tab. 1), built from MIMIC-CXR [10,12,13] images, reports, and Chest ImaGenome scene graphs [24,25]. The dataset follows the official MIMIC-CXR split, excluding the MS-CXR-T [2,3] test set. We develop a report processing pipeline (Fig. 1) to extract structured information. Below, we detail the sub-datasets.

1. Clean image-report pairs. To align VL representations, we pair MIMIC-CXR images with their radiological reports. We retain only AP/PA view scans and use Gemini to filter reports, keeping only view information, FINDINGS, and IMPRESSION sections while removing comparative descriptions. Using this sub-dataset, we can pretrain CoCa in the training stage 1 to align a CXR image with its corresponding image description.

2. Image pairs & filtered reports. This subset contains samples with explicit comparison. Each sample consists of an image, its most recent prior image, and its FINDINGS and IMPRESSION sections as textual descriptions.

We make sure each report in this sub-dataset contains both description of the current image and the comparison between the current and prior images.

3. Image pairs & comparative descriptions. Building on subset 2, the subset 3 only includes sentences explicitly describing progression, extracted via Gemini. This subset can be regarded as a strong supervision for the model to learn the correspondence between an image pair and its progression. To augment the data, we reverse image pairs and modify text descriptions accordingly (*e.g.*, "improved pneumonia" → "worsened pneumonia").

4. Image pair & abnormal organs. Using Chest ImaGenome, we extract structured comparative descriptions, providing localized abnormality progression. The format follows: "[condition] [progress] at [organ], coordinates for current image is $[x_{cur,1}, x_{cur,2}, y_{cur,1}, y_{cur,2}]$, coordinates for previous image is $[x_{prior,1}, x_{prior,2}, y_{prior,1}, y_{prior,2}]$." This sub-dataset provides the model with regional annotation as a finer-grained supervision. We also reverse image pairs and adjust descriptions accordingly to augment the dataset.

3 CoCa-CXR

3.1 Contrastive Captioners for CXR Image Pair Understanding

CoCa [28] (top left in Fig. 1) encodes images and text into latent representations using a Vision Transformer (ViT) [8] and a unimodal text encoder. A transformer text decoder then cross-attends to image features to generate captions. The model is trained with a multi-modal contrastive loss:

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(y_i^\top x_i/\tau)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\tau)}\right)$$

where x_i and y_i are normalized image and text embeddings, N is the batch size, and τ is a temperature parameter. CoCa also learns fine-grained representations through its text decoder with an autoregressive captioning objective:

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x).$$

The final training objective combines both losses:

$$\mathcal{L}_{\text{CoCa}} = \mathcal{L}_{\text{Con}} + \lambda\mathcal{L}_{\text{Cap}}.$$

To adapt CoCa for modeling CXR image pairs and their temporal differences, we use its ViT encoder to extract embeddings for the current (z^c) and prior (z^p) images, where $z \in \mathbb{R}^{N\times d}$, with N as the sequence length and d as the feature dimension. The regional cross-attention module (detailed in Sec. 3.3) processes z^c as the main input and z^p as the cross-attention input, producing an output embedding $z^o \in \mathbb{R}^{N\times d}$. We then concatenate these embeddings to form the final visual token sequence:

$$z^{concat} = concat(z^c, z^p, z^o) \in \mathbb{R}^{3N\times d}.$$

This modified architecture, CoCa-CXR, enables the model to capture both individual image semantics and temporal progression between CXR image pairs.

3.2 Three-Stage Training of CoCa-CXR

Training CoCa-CXR in a single stage proved challenging (see Sec. 4.2 ablation study), so we adopt a three-stage training strategy (Fig. 1). The stage 1 training is performed on sub-dataset 1, and the stage 2 and 3 use all the 4 sub-datasets.

Stage 1: CoCa pretraining. We first train CoCa on the clean image-report pairs to establish a general alignment between language and visual patterns.

Stage 2: Regional attention pretraining. Next, we train the regional cross-attention module using image pairs. This step is crucial because the module is randomly initialized, unlike the pretrained backbone. Prior work [5,15] shows that prioritizing less transferable parameters improves performance.

Stage 3: Supervised fine-tuning. Finally, we fine-tune the regional cross-attention module and multi-modal text decoder using all four sub-datasets, refining the models ability to capture temporal progression and generate reports.

3.3 Regional Cross-Attention Module

The regional cross-attention module is a Transformer block [22] that processes embeddings from the current (z^c) and prior (z^p) CXR images. First, a self-attention layer refines z^c:

$$z^{c'} = \text{SelfAttention}(z^c) = \text{softmax}\left(\frac{Q_c K_c^T}{\sqrt{d}}\right) V_c,$$

where $Q_c = z^c W_Q$, $K_c = z^c W_K$, and $V_c = z^c W_V$ are the query, key, and value projections, with learnable weights W_Q, W_K, W_V.

Next, a cross-attention layer with **regional masking** (Fig. 1, bottom right) extracts localized differences between images. The query $Q'_i = z^{c'} W'_Q$ from the current image attends to a restricted set of key-value pairs from the prior image:

$$\text{CrossAttention}_i = \text{softmax}\left(\frac{Q'_i {K'}^T_{\text{region}(i)}}{\sqrt{d}}\right) V'_{\text{region}(i)},$$

where $K'_{\text{region}(i)} = K' \odot M_{\text{region}(i)}$ and $V'_{\text{region}(i)} = V' \odot M_{\text{region}(i)}$ are masked key-value pairs within a local window around Q'_i. The mask $M_{\text{region}(i)}$ selects relevant regions in the prior image, where $region(i)$ refers to a local square window around position i in the image token grid. Although different abnormalities may have overlapped regions, our model applies a consistent spatial restriction to encourage localized comparison, regardless of the disease category.

The final output sequence $z^o \in \mathbb{R}^{N \times d}$ is obtained by passing $\{\text{CrossAttention}_i\}_{i=1}^N$ through a feed-forward network, followed by skip connections and normalization.

Table 2: Comparison on MS-CXR-T temporal image classification dataset (repeated for 4 random seeds). We report the macro-accuracy (%) across the three progression classes (worsened, unchanged, improved) for each condition following BioViL-T. BioViL-T does not explicitly discuss its validation set. For a complete comparison, we report both MS-CXR-T validation and testing performance.

Method	Consolidation	Pleural effusion	Pneumonia	Pneumothorax	Edema	Avg
CNN + Transformer [2]	44.0 ± 2.0	61.3 ± 1.6	45.1 ± 3.5	31.5 ± 3.1	65.5 ± 1.1	49.5
CheXRelNet [14]	47	47	47	36	49	45.2
BioViL [4]	56.0 ± 1.5	63.0 ± 0.9	60.2 ± 0.7	42.5 ± 2.7	67.5 ± 0.9	57.8
BioViL-T [2]	61.1 ± 2.4	67.0 ± 0.8	61.9 ± 1.9	42.6 ± 1.6	68.5 ± 0.8	60.2
Med-ST [26]	60.6 ± 1.2	67.4 ± 0.3	58.5 ± 1.5	65.0 ± 0.3	54.2 ± 0.8	61.1
CoCa-CXR (val.)	70.4 ± 0.5	69.6 ± 1.7	61.4 ± 1.6	72.8 ± 1.1	71.8 ± 0.3	69.2
CoCa-CXR (test)	69.6 ± 2.5	68.1 ± 1.5	56.4 ± 0.8	59.3 ± 2.6	71.8 ± 0.8	65.0

4 Experiments & Results

4.1 Experimental Setting

For all training stages, CXR images are padded to square, resized to 768×768 pixels, and normalized to [0,1] without additional augmentations. The CoCa image encoder extracts $48^2 = 2304$ visual tokens per image, which are processed by the regional cross-attention module using an 11^2 masking window. The resulting sequence is downsampled via 2D average pooling to $16^2 = 256$ tokens before entering the multi-modal text decoder. We train with AdamW, using a learning rate of 2×10^{-5} for stages 1 and 3, and 10^{-4} for stage 2. The model is optimized with a batch size of $N = 64$ for 20k, 10k, and 30k iterations in each stage, respectively. The sub-dataset ratio for the last two stages is set to $0.2 : 0.25 : 0.25 : 0.3$.

4.2 Evalutation of CoCa-CXR

Results on Temporal Classification. On the MS-CXR-T dataset [2], CoCa-CXR predicts condition progression between two images using the prompt "[condition] is ", selecting the most probable next token from {"worsened", "unchanged", "improved"}. As shown in Tab. 2, CoCa-CXR outperforms previous SOTA models on both validation and test sets. It surpasses BioVil-T in four out of five conditions, achieving an average test accuracy of 65.0%, which is 4.8% higher than BioVil-T.

Results on Report Generation. Tab. 3 presents report generation results on the MIMIC-CXR dataset [13]. CoCa-CXR predicts both FINDINGS and IMPRESSION sections, a more challenging task than FINDINGS alone [2]. When generating only descriptions, it achieves a RadGraph F1 score of 24.2%, on par with large-scale models like Med-Gemini [27]. Notably, PaliGemma-2 [18] incorporates both images and the indication section as inputs, whereas CoCa-CXR

Table 3: CXR report generation on the MIMIC-CXR dataset with metrics (%) sourced from published research. We show both CoCa-CXR results with description only and description + comparison.

Method	Section	RadGraph F1	BLEU4	Rouge-L
CXR-RePaiR [9]	Findings	9.1	2.1	14.3
M^2 Transformer [16]	Findings	22.0	11.4	-
Med-PaLM M, 12B [21]	Findings	25.2	10.4	26.2
CvT-21DistillGPT2 [17]	Findings + Impression	15.4	12.4	28.5
Flamingo-CXR [19]	Findings + Impression	20.5	10.1	29.7
Med-Gemini-2D [27]	Findings + Impression	24.4	20.5	28.3
CoCa-CXR (des. only)	Findings + Impression	24.2	18.6	27.8
CoCa-CXR (des. + comp.)	Findings + Impression	23.7	18.7	27.5

Table 4: Ablation study on dataset construction, attention module, and the model training scheme. We report testing accuracy (%) on MS-CXR-T dataset.

	Ablation	Con.	Pl. Eff.	Pneumon.	Pneumoth.	Edema	Avg
	CoCa-CXR	69.6	68.1	56.4	59.3	71.8	65.0
Dataset	w/o Cleaning single image description	64.8	70.0	58.5	56.4	68.4	63.6
Dataset	w/o Filtering comparing pairs	65.2	71.2	54.9	57.9	70.9	64.0
Dataset	w/o Comparison-only description	59.8	65.3	60.6	47.7	69.9	60.7
Dataset	w/o Abnormal organs & coordinates	54.2	69.1	59.9	53.2	65.9	60.5
Model	w/o Regional cross-attention	58.8	70.5	58.8	47.2	69.6	61.0
Model	w/o Contrastive learning	57.4	68.5	49.8	45.0	70.7	58.3
Model	w/o Stage 2 pretraining	61.3	69.1	55.8	52.4	67.9	61.3
Model	w/o Stage 1 and 2 pretraining	58.5	65.5	58.9	46.2	62.9	58.4

and other models in Tab. 3 rely solely on images; hence, we exclude its result for fair comparison. When also generating comparison descriptions, CoCa-CXR attains a RadGraph F1 score of 23.7%. Although this is slightly lower than description-only generation, our hypothesis is that, describing progression adds another dimension to report generation tasks, and with the current accuracy-level (65%) on temporal classification, one can obtain a similar level of benefit on Radgraph F1 by simply omitting comparisons in the generated reports. But still, as shown in Fig. 2, this is step forward in incorporating this progression dimension and towards real world application.

Ablation Study. To assess the impact of CoCa-CXRs components, we perform ablation studies on dataset construction and model training (Tab. 4). For temporal classification, we find that the comparison-only descriptions in sub-dataset 3 and the abnormal organ annotations with coordinates in sub-dataset 4 are crucial. The proposed regional cross-attention module improves average accuracy from 61.0% to 65.0%, demonstrating its effectiveness in capturing temporal differences. Additionally, contrastive loss enhances representation learning for identifying image variations. Finally, both stage 1 and stage 2 pretraininginitial-

Ground Truth

(Image 2 as prior) Findings: Support and monitoring devices are unchanged in position, and **cardiomediastinal contours are similar**. **Interval worsening of pulmonary edema** as well as slight **increase in size of moderate bilateral pleural effusions**. Otherwise, no relevant short interval change.

Prediction

(Image 2 as prior) Findings: The et tube tip is approximately 6 cm above the carina. The right internal jugular line tip is at the level of superior svc. **Heart size and mediastinum are stable**. There is **interval progression of pulmonary edema**. Bilateral pleural effusions are noted, unchanged.

(Image 1 as prior) Findings: The et tube tip is approximately 6 cm above the carina. The ng tube tip is in the stomach. The right internal jugular line tip is at the level of superior svc. **Heart size and mediastinum are unchanged**. There is **interval improvement of pulmonary edema.**

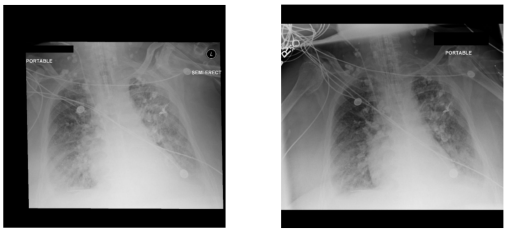

Image 1 Image 2

Fig. 2: Report generation of CoCa-CXR on MIMIC-CXR validation set. If we swap the order of the image pair, the comparison prediction changes accordingly.

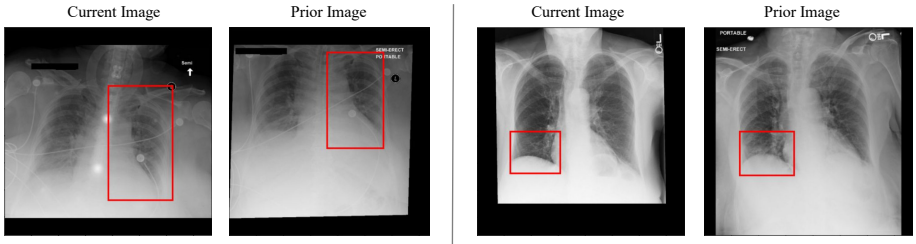

Fig. 3: Visualization of the text-based condition progression detection.

izing encoders and the regional cross-attention moduleare essential, highlighting the importance of our three-stage training strategy.

Visualization. We visualize the learned capability of CoCa-CXR through its generated report and abnormality detection. The Fig. 2 shows that the generated report can correctly describe the image content and the change from prior to current image. After swapping the order of two images, the prediction also reverse. In Fig. 3, CoCa-CXR predicts the condition, progression, and the coordinates in two images demonstrating the model's capability of localizing abnormal organs. Specifically, the Intersection over Union (IoU) for Left lower lung is 0.589 on the validation set for sub-dataset 4. A full performance breakdown on 10 pulmonary structures is provided in the supplementary material. These

results highlight the role of vision-language alignment pretraining and regional cross-attention in capturing localized CXR patterns.

5 Conclusion

This work demonstrates how leveraging an LLM to curate condensed temporal information (CXR-4) enhances the training of a temporally aware model, CoCa-CXR. We introduce a regional cross-attention module to improve longitudinal CXR analysis by guiding attention across time. CoCa-CXR surpasses previous SOTA in temporal classification by incorporating explicit comparison supervision and regional attention within a three-stage training framework. It accurately predicts disease progression and generates reports with RadGraph F1 scores comparable to leading models.

6 Acknowledgment

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Faruk Ahmed for the feedback and valuable insights.

7 Disclosure of Interests.

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., Meissen, F., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
2. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
3. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Coelho de Castro, D., Boecking, B., Sharma, H., Bouzid, K., Schwaighofer, A., Wetscherek, M.T., Richardson, H., Naumann, T., Alvarez Valle, J., Oktay, O.: Ms-cxr-t: Learning to exploit temporal structure for biomedical vision-language processing (version 1.0.0). PhysioNet (2023), https://doi.org/10.13026/pg10-j984
4. 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)
5. Chen, Y., Liu, L., Li, J., Jiang, H., Ding, C., Zhou, Z.: Metalr: Meta-tuning of learning rates for transfer learning in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 706–716. Springer (2023)

6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Cho, Y., Kim, T., Shin, H., Cho, S., Shin, D.: Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. In: Medical Imaging with Deep Learning (2024)
8. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health. pp. 209–219. PMLR (2021)
10. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)
11. 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)
12. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database (version 2.0.0). PhysioNet (2019), https://doi.org/10.13026/C2JT1Q
13. 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)
14. 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)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
16. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)
17. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
18. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al.: Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555 (2024)
19. Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Singhal, K., Azizi, S., Tu, T., et al.: Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. arXiv preprint arXiv:2311.18260 (2023)
20. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

21. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. NEJM AI **1**(3), AIoa2300138 (2024)
22. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
23. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200. Springer (2024)
24. Wu, J., Agu, N., Lourentzou, I., Sharma, A., Paguio, J., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Syeda-Mahmood, T., Moradi, M.: Chest imagenome dataset (version 1.0.0). PhysioNet (2021), https://doi.org/10.13026/wv01-y230
25. 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)
26. Yang, J., Su, B., Zhao, W.X., Wen, J.R.: Unlocking the power of spatial and temporal information in medical multimodal pre-training. arXiv preprint arXiv:2405.19654 (2024)
27. Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al.: Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024)
28. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)