HERGen: Elevating Radiology Report Generation with Longitudinal Data

Fuying Wang[1], Shenghui Du[1], and Lequan Yu[1]

The University of Hong Kong, Hong Kong
{fuyingw@connect.,shenghui@connect.,lqyu@}hku.hk

Abstract. Radiology reports provide detailed descriptions of medical imaging integrated with patients' medical histories, while report writing is traditionally labor-intensive, increasing radiologists' workload and the risk of diagnostic errors. Recent efforts in automating this process seek to mitigate these issues by enhancing accuracy and clinical efficiency. However, existing automated approaches are based on a single timestamp and often neglect the critical temporal aspect of patients' imaging histories, which is essential for accurate longitudinal analysis. To address this gap, we propose a novel **H**istory **E**nhanced Radiology **R**eport **Gen**eration (**HERGen**) framework that employs a group causal transformer to efficiently integrate longitudinal data across patient visits. Our approach not only allows for comprehensive analysis of varied historical data but also improves the quality of generated reports through an auxiliary contrastive objective that aligns image sequences with their corresponding reports. More importantly, we introduce a curriculum learning-based strategy to adeptly handle the inherent complexity of longitudinal radiology data and thus stabilize the optimization of our framework. The extensive evaluations across three datasets demonstrate that our framework surpasses existing methods in generating accurate radiology reports and effectively predicting disease progression from medical images.

Keywords: Radiology Report Generation · Longitudinal Study · Vision-Language Learning

1 Introduction

Chest X-rays are a cornerstone in diagnosing thoracic conditions, including pneumonia and lung cancer [19,35]. Given a chest X-ray, radiologists will meticulously examine each anatomical section in the X-ray and document their observations with detailed text descriptions. The generated report is crucial to diagnose diseases (*e.g.*, lung cancer, scoliosis) and assess the position of the treatment devices (*e.g.*, tracheostomy tubes, pacemakers). Particularly, when prior images are available, radiologists commonly compare the clinical findings of the current scan with prior scans to assess the evolution of disease over time, which is essential in regular clinical evaluations. However, the high volume of chest X-rays overwhelms radiologists, exacerbating the impact of the global shortfall in this workforce [8, 37].

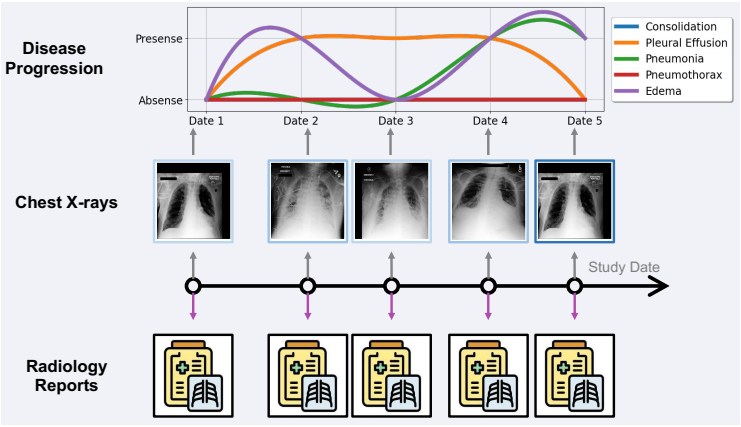

Fig. 1: Overview of our HERGen for radiology report generation: Our model processes longitudinal data for each patient and utilizes the comprehensive historical information within these longitudinal data to generate robust and precise radiology reports.

Automated chest X-ray report generation has emerged as a key research area, aiming to ease radiologists' workload and improve patient care [43]. Mainstream approaches focus on improving clinical accuracy and completeness of individual reports [9, 10, 27, 42], often overlooking the chronological consistency in longitudinal imaging. Modeling such inherent temporal information in chest X-rays has shown to be crucial for generating precise radiology reports [5, 20, 39, 59]. Some recent studies integrate prior images for temporal representation and enhance report generation [5, 39]. However, they are limited to the use of only one prior image for the current report, failing to capture high-level disease progression evident across a patient's history. This highlights the need for a framework that learns accurate representations from both study-level and patient-level images, thereby producing reports closely aligned with radiologists' analyses.

In this paper, we propose a novel **H**istory **E**nhanced radiology **R**eport **Gen**eration framework (HERGen) to effectively capture the temporal information of longitudinal data for generating comprehensive and temporally coherent radiology reports, as shown in Fig. 1. The key part is a causal transformer model, which treats all visual tokens from the same image as a group and uses a group causal attention mechanism to handle it. Viewing all visual tokens of each patient as a sequence, this mechanism groups visual tokens from the same image, facilitating intra-image interactions of visual tokens and inter-image interactions of tokens only across previous studies. Notably, it treats each patient's X-ray series as a distinct sequence, adeptly handling the variability in the number of longitudinal images per patient. Moreover, we further refine the model's capability to chart disease progression through a cross-modal contrastive learning objective, ensuring the alignment of longitudinal visual representations with their narrative reports. Due to the inherent complexity of longitudinal data, it is non-trivial to optimize the whole framework. We thereby introduce a new curriculum learning-

based optimization strategy in three progressive steps to enhance and stabilize the learning process of our framework. The model is first trained to generate radiology reports for individual images and then, we employ the auxiliary contrastive alignment module to optimize the latent space. After that, the entire framework is trained with the integration of a temporal aggregation module, enabling it to learn from the patient's historical information. Extensive experimental results on radiology report generation and temporal medical image classification tasks demonstrate the superiority of our framework in generating accurate radiology reports and effectively predicting the disease progression from medical images. The source code is available at `https://github.com/fuying-wang/HERGen`.

2 Related Work

Automated Report Generation. Radiology report generation, inspired by image captioning techniques [11, 46, 54, 56], face unique challenges due to the complexity and variability in radiology reports [1, 9, 10, 18, 23, 25, 49, 55]. Initial approaches, primarily based on CNN-RNN [17, 18, 49, 57], have evolved with the adoption of transformer [44]. Recent advancements include memory-driven transformers for enhanced cross-modal interactions [9,10], alignment of visual features with disease tags [55], and contrastive methods for anomaly detection [25]. Integration of knowledge graphs [23, 57], warm starting strategies [27], and interactive frameworks for region-specific reports [42] have also been explored. However, these methods often treat X-rays and reports as independent entities, overlooking the temporal aspects inherent in various radiology modalities.

Longitudinal Chest X-ray Representations. Radiology studies, inherently chronological, are crucial for accurate reporting, yet the temporal dimension is often under-addressed in research. [34] indirectly acknowledged the importance of sequential context by proposing a method to reduce language model hallucinations. [5] introduced a self-supervised framework capturing the longitudinal evolution of chest X-ray findings. Similarly, [59] developed a cross-attention-based multi-modal fusion framework utilizing patient record chronology to enhance report pre-filling. [20] employed graph attention networks [45] for an anatomy-aware approach to tracking disease progression in longitudinal CXR data. [39] used Faster R-CNN [36] to project longitudinal studies into a composite representation highlighting anatomical changes over time. However, most of these methods primarily focus on learning representations rather than generating reports. Furthermore, these methods often treat two consecutive image-text pairs, lacking flexibility for varying patient history lengths and are limited in capturing the complex progression of diseases.

Biomedical Vision-language Pretraining. Radiology reports, paired with chest X-rays, offer rich labels for learning visual representations. Building on the CLIP framework [32], [7, 15, 47, 58] demonstrate the efficacy of self-supervised vision-language pretraining in biomedical imaging tasks. Particularly, [58] use a contrastive objective [29] for modality alignment, [15] focus on local alignment

for detailed feature learning, and [7] develop CXR-BERT, employing masked language modeling for enhanced image feature learning from radiology language. Furthermore, [5] adapt BioViL [7] for longitudinal analysis in radiology, improving temporal aspects in report generation and classification tasks. Our work further explores the application of vision-language pretraining on longitudinal data, aiming to effectively capture disease progression in patient records.

3 Method

3.1 Problem Formulation

The overall framework of the proposed method is shown in Fig. 2. We analyze a dataset comprising chest X-rays from M patients, denoted as $\{\mathcal{I}_i\}_{i=1,2,...,M}$, where $\mathcal{I}_i = \{\mathbf{I}_j^{(i)}\}_{j=1,2,...,N_i}$ represents the set of X-rays for the i-th patient and N_i is the number of studies (visits). For each patient i, their X-rays are chronologically ordered based on their associated study dates $\mathcal{T}_i = \{\mathbf{T}_j^{(i)}\}_{j=1,2,...,N_i}$. The objective of our method is to generate a set of radiology reports $\{\hat{\mathcal{R}}_i\}_{i=1,2,...,M}$ for each patient, aiming to closely approximate the ground truth reports $\{\mathcal{R}_i\}_{i=1,2,...,M}$. In the following, we use i and j to index patient and study respectively.

3.2 History-enhanced Report Generation

Extract Representations of Single Images. In our approach, each X-ray image $\mathbf{I}_j^{(i)} \in \mathbb{R}^{C\times W\times H}$ is first encoded into a feature representation $\mathbf{P}_j^{(i)} \in \mathbb{R}^{S\times F}$ with an image encoder. Here, C, W, and H denote the number of channels, width, and height of the image, respectively, while S and F represent the number of visual tokens and the feature dimension per token. Following CvT-212DistilGPT2 [27], we utilize the CvT architecture [52], pretrained on ImageNet-21K, as our image encoder, while our framework can take other encoder backbones. To tailor the dimensions of S and F to our requirements, we introduce an encoder projection layer E_{proj}. This layer comprises a 1×1 convolution layer followed by a linear projection layer, transforming each $\mathbf{P}_j^{(i)}$ into a more compact visual representation $\mathbf{V}_j^{(i)} \in \mathbb{R}^{S'\times F'}$, where S' and F' denote the adjusted number of visual tokens and their new dimensionality, respectively.

Sequential Date-aware Temporal Embedding. Temporal embeddings are especially critical for our group causal transformer to learn longitudinal information. Standard positional embeddings typically assume equidistant intervals between tokens, an assumption that is not applicable in our context due to the varying time gaps between consecutive chest X-rays. For example, the clinical progression captured in X-rays taken a month apart is significantly different from that in X-rays taken a year apart. To tackle this challenge, we introduce study date-aware positional embeddings, $\mathbf{p}_j^{(i)} \in \mathbb{R}^{S^*\times F'}$ for each study. These embeddings are conditioned on the study dates $\mathbf{T}_j^{(i)}$, offering a more precise representation of the temporal intervals between X-rays. In detail, we first calculate

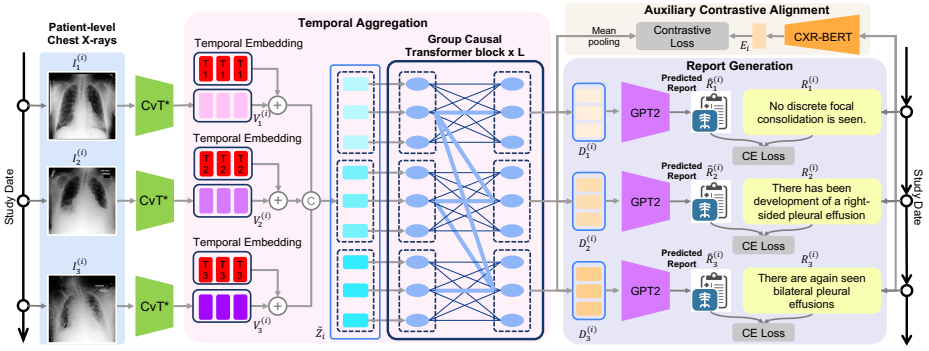

Fig. 2: History **E**nhanced Radiology **R**eport **Gen**eration (HERGen): the framework processes patient-level chest X-rays using the CvT* (CvT combined with the encoder projection layer), which then aggregates temporal information through a group causal transformer. Subsequently, GPT2 serves as the decoder for predicting the radiology report, which was optimized by a cross-entropy (CE) loss. Additionally, an auxiliary contrastive alignment module is employed to enhance the alignment of the latent spaces between image and text modalities, thereby producing more consistent reports. Note that in the group causal transformer block, thick lines represent image-level interactions, while thin lines indicate token-level interactions.

the relative study date for each X-ray image as $\mathbf{T}'^{(i)}_j = \mathbf{T}^{(i)}_j - \mathbf{T}^{(i)}_0$. Then, we identify the maximum relative study date in the training set and create a learnable embedding vocabulary of the corresponding length. Each temporal embedding is defined as: $\mathbf{p}^{(i)}_j = \text{Embedding}(\mathbf{T}'^{(i)}_j) \in \mathbb{R}^{1\times F'}$. The visual token embeddings $\mathbf{V}^{(i)}_j$ are then added with the temporal embedding $\mathbf{p}^{(i)}_j$ to form $\mathbf{Z}^{(i)}_j = \mathbf{V}^{(i)}_j + \mathbf{p}^{(i)}_j$, where $\mathbf{Z}^{(i)}_j \in \mathbb{R}^{S'\times F'}$. Finally, we concatenate all visual token embeddings for each patient to create a patient-level sequence $\tilde{\mathbf{Z}}_i = \text{Concat}([\{\mathbf{Z}^{(i)}_j\}_{j=1,2,\ldots,N_i}])$ where $\tilde{\mathbf{Z}}_i \in \mathbb{R}^{S^{(i)}\times F'}$ and $S^{(i)} = N_i \times S'$, which is then fed into the group causal transformer for temporal aggregation.

Group Causal Transformer. Our group causal transformer comprises L group causal blocks, designed to aggregate longitudinal information from patient data. In block l, for every visual token (indexed by p), we first compute the query, key, and value vectors from its preceding block's representation $\mathbf{z}^{(l-1)}_{(p)} \in \mathbb{R}^{F'}$ as:

$$\begin{aligned}
\mathbf{q}^{(l,a)}_{(p)} &= \mathbf{W}^{(l,a)}_Q \text{LN}(\mathbf{z}^{(l-1)}_{(p)}) \in \mathbb{R}^{D_h}, \\
\mathbf{k}^{(l,a)}_{(p)} &= \mathbf{W}^{(l,a)}_K \text{LN}(\mathbf{z}^{(l-1)}_{(p)}) \in \mathbb{R}^{D_h}, \\
\mathbf{v}^{(l,a)}_{(p)} &= \mathbf{W}^{(l,a)}_V \text{LN}(\mathbf{z}^{(l-1)}_{(p)}) \in \mathbb{R}^{D_h}.
\end{aligned} \tag{1}$$

Here, LN denotes LayerNorm and $a = 1, \ldots, A$ indexes the A attention heads with the latent dimensionality for each head being $D_h = F'/A$. The initial representation $\mathbf{z}^{(0)}$ corresponds to the input sequence $\tilde{\mathbf{Z}}_i$. $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$

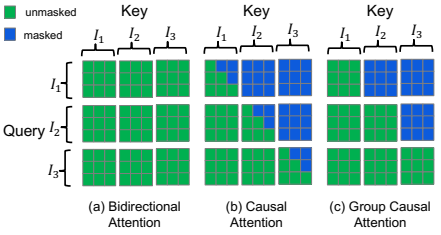

Fig. 3: Comparison between (a) bidirectional attention, (b) causal attention and (c) our group causal attention.

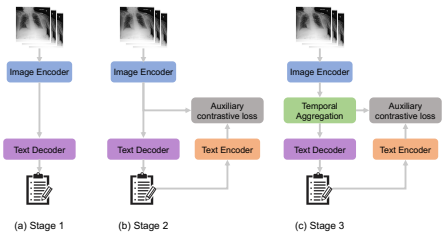

Fig. 4: Illustration of the proposed curriculum training strategy.

are learnable matrices. For simplicity, we have omitted the patient index i in our notation. Then, the process for computing dot-product self-attention weights, along with subsequent steps, is defined as follows:

$$\begin{aligned} \alpha_{(p)}^{(l,a)} &= \mathrm{SM}\left(\frac{\mathbf{q}_{(p)}^{(l,a)}}{D_h} \cdot [\mathbf{k}_{(p')}^{(l,a)}]_{p'\in 0,\ldots,S^{(i)}-1}\right), \\ \mathbf{s}(p)^{(l,a)} &= \sum_{p'} \mathbf{M}(p')\alpha(p)(p')^{(l,a)}\mathbf{v}(p')^{(l,a)}, \end{aligned} \tag{2}$$

where $\alpha \in \mathbb{R}^{S^{(i)}}$. The group causal attention matrix $\mathbf{M}(p')$, as illustrated in Fig. 3, differs fundamentally from the bidirectional attention used in BERT [12] and the causal attention in GPT [33]. It ensures that each visual token within an image not only interacts with others in the same image but also with tokens from preceding images.

This design reflects our intention to make the transformer cognizant of the temporal sequence in radiological data, a crucial aspect for accurately capturing disease progression over time. Subsequently, we perform concatenation followed by a Multi-Layer Perceptron (MLP) with residual connections to get the output. This can be mathematically represented as:

$$\begin{aligned} \mathbf{z'}_{(p)}^{(l)} &= \mathbf{W}_O[\mathbf{s}_{(p)}^{(l,1)}...\mathbf{s}_{(p)}^{(l,A)}] + \mathbf{z}_{(p)}^{(l-1)}, \\ \mathbf{z}_{(p)}^{(l)} &= \mathrm{MLP}(\mathrm{LN}(\mathbf{z'}_{(p)}^{(l)})) + \mathbf{z'}_{(p)}^{(l)}, \end{aligned} \tag{3}$$

where $\mathbf{W}_O$ is a learnable matrix. Then, the output sequence of the group causal transformer, denoted as $\mathbf{z}^{(L)} \in \mathbb{R}^{S^{(i)}\times F'}$, is split into a series of representations of studies $\{\mathbf{D}_j^{(i)}\}_{j=1,\ldots,N_i}$ and $\mathbf{D}_j^{(i)} \in \mathbb{R}^{S'\times F'}$.

Report Generation and Auxiliary Contrastive Alignment. Each temporally aggregated visual representation $\mathbf{D}_j^{(i)}$ is input into a text decoding module for generating radiology reports. Note that We chose GPT-2 as the text decoder following [27], which shows DistilGPT2 [38] outperforms other alternatives like ClinicalBERT [2], PubMedBERT [13], and SciBERT [6]. We minimize a cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ to ensure predicted reports are close to ground truth reports.

To improve the coherence of generated reports, we introduce an auxiliary contrastive alignment module. This module is designed to align the distributions of the visual and textual modalities, thereby enhancing the model's overall performance. Initially, for each visual token embedding $\mathbf{D}_j^{(i)} \in \mathbb{R}^{S' \times F'}$, we perform a mean pooling operation along the first dimension and it results in a global representation of the entire image, denoted as $\tilde{\mathbf{D}}_j^{(i)} \in \mathbb{R}^{F'}$. Then, we use a text encoder to encode each report $\mathbf{R}_j^{(i)}$ into a representation $\mathbf{E}_j^{(i)}$. Subsequently, we concatenate all visual and text embeddings within the same minibatch to form a combined set: $\{\tilde{\mathbf{D}}_s\}_{s=1,2,\ldots,N_B}$ and $\{\mathbf{E}_s\}_{s=1,2,\ldots,N_B}$, respectively. Note that N_B represents the total number of studies within the mini-batch. The contrastive loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{Cont}} = \sum_{r=1}^{N_B} \frac{1}{2N_B} (&-\log \frac{\exp(\text{sim}(\tilde{\mathbf{D}}_r, \mathbf{E}_r)/\tau)}{\sum_{s'=1}^{N_B} \exp(\text{sim}(\tilde{\mathbf{D}}_r, \mathbf{E}_{s'})/\tau)} \\ &- \log \frac{\exp(\text{sim}(\mathbf{E}_r, \tilde{\mathbf{D}}_r)/\tau)}{\sum_{s'=1}^{N_B} \exp(\text{sim}(\mathbf{E}_r, \tilde{\mathbf{D}}_{s'})/\tau)})\end{aligned} \tag{4}$$

where τ is the temperature hyperparameter.

Learning Objectives. Finally, our model is optimized by jointly minimizing these two objectives:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{Cont}}. \tag{5}$$

Here, λ is a hyperparameter used to balance these two losses. Based on empirical studies, we set λ as 1.0. The ablation results of the hyperparameter r can be found in the Supplementary Material.

3.3 Curriculum Training

As shown in Fig. 4, we introduce a curriculum learning strategy, unfolding in three stages to progressively enhance our model's performance:

- **Stage 1: Encoder-Decoder Report Generation**: Initially, reports are generated using an encoder-decoder architecture trained on individual chest X-ray image-text pairs. This foundational step focuses solely on static data without temporal context.
- **Stage 2: Alignment Refinement with Text Encoder**: Subsequently, a text encoder is incorporated, utilizing contrastive learning to refine the alignment between the visual and textual data.
- **Stage 3: Temporal Information Learning**:The final stage expands the model's capability to a longitudinal perspective. Here, we integrate the group causal transformer to process sequences of chest X-rays, thereby incorporating temporal information into the report generation.

These stages collectively develop a robust and comprehensive model, which is then systematically evaluated to assess its effectiveness in generating accurate and contextually relevant radiology reports.

4 Experiments

4.1 Experimental Setup

Dataset and Preprocessing. We evaluate the performance of our model on two clinical tasks: *radiology report generation* and *temporal medical image classification*. The used datasets are as follows:

- **MIMIC-CXR:** We utilize the MIMIC-CXR dataset [19], which originally comprises 377,110 chest X-ray images and 227,835 reports, to evaluate our model. Aligning with previous work [9, 10, 27], we adopt the official split of the MIMIC-CXR dataset in our experiment. However, the original dataset includes multiple lateral images, which could introduce inconsistency in longitudinal analyses. Additionally, we observed duplicate images within the same study, bringing noise to patient-level progression analysis. Therefore, we meticulously curated the dataset by removing lateral images and duplicates within studies for each train/validation/test set, resulting in a preprocessed dataset consisting of 145,471 pairs for training, 1,151 for validation, and 2,210 for testing. We then follow [27] to preprocess images and reports. Specifically, we resize all images to 384×384 while preserving aspect ratios. Report preprocessing involved truncating to 60 words, converting to lowercase, removing special characters, and replacing infrequent terms with placeholders. Crucially, we organized the image-report pairs chronologically based on the "StudyDate" metadata, preserving temporal integrity for analyzing each patient's radiological history. Further details on dataset curation and preprocessing are available in the Supplementary Material. Note that we re-run the publicly released code of compared methods on our curated MIMIC-CXR dataset to ensure a fair comparison.
- **Longitudinal MIMIC-CXR:** We further devise the Longitudinal MIMIC-CXR dataset, derived from the preprocessed MIMIC-CXR-JPG dataset, to assess our model's capability in generating temporally coherent reports, following [59]. This subset includes only patients with at least two consecutive visits. It is worth noting that the training, validation, and test splits of the Longitudinal-MIMIC dataset correspond to the official divisions of the MIMIC-CXR dataset.
- **MS-CXR-T:** We also assess our model's capacity for capturing temporal information using the MS-CXR-T dataset [4]. This dataset consists of 1326 multi-image frontal chest X-rays, each annotated with one of five findings. For each finding, there are three possible states reflecting disease progression: "Improving," "Stable," and "Worsening".

Radiology Report Generation. We evaluate the performance of our method in radiology report generation on both the MIMIC-CXR dataset and the Longitudinal MIMIC-CXR dataset. We compare HERGen with 7 state-of-the-art (SOTA) radiology report generation models, including $\mathcal{M}^2$Transformer [11], R2Gen [10], R2GenCMN [9], $\mathcal{M}^2$TR.PROGRESSIVE [28], XProNet [48], CvT-212DistilGPT2 [27] and DCL [21]. To ensure a fair comparison, we rerun the

publicly released code of these methods on our curated MIMIC-CXR dataset. Note that 5 SOTA models (PPKER [23], ContrastiveAttention [25], AlignTransformer [55], KIUT [16], and METransformer [50]) lack publicly available source code, thus results are cited from their original papers for reference. However, we note that these can't directly compared to our results due to our additional dataset preprocessing. Additionally, the results in the RGRG [42] paper, employing the Chest ImaGenome [53] split instead of the official MIMIC-CXR split, are also for reference only but not directly comparable. On the Longitudinal MIMIC-CXR dataset, we compare our model with both single-image based baselines, i.e., R2Gen [10], R2CMN [9], CvT-212DistilGPT2 [27], etc. and longitudinal image-based baseline, i.e., Prefilling [59].

Temporal Image Classification. The temporal image classification task is evaluated on the MS-CXR-T dataset [4]. This evaluation serves as an additional task to assess how well our model can understand and process disease progression in medical images. We compare our approach with both temporal image-based vision language pretraining methods (*e.g.*, BioViL-T) and single image pretraining methods (*e.g.*, BioViL). More information about this experiment is available in the Supplementary Material.

Evaluation Metrics. In line with previous studies [10, 27, 31, 42], we employed a combination of Natural Language Generation (NLG) and Clinical Efficiency (CE) metrics to evaluate our report generation performance. For NLG, we used established metrics including BLEU-n [30], which measures n-gram overlap, METEOR [3], that accounts for recall through an F_β score, ROUGE-L [22], based on the longest common subsequence. Recognizing that NLG metrics may not fully reflect clinical accuracy, we further integrated CE metrics following previous work [9, 10, 16, 50]. Specifically, we apply CheXbert [40] to label the generated reports into 14 categories (related to thoracic diseases and support devices), and then compute precision, recall, and F1 scores against ground truths. The macro-averaged results over 14 classes are reported, given the susceptibility of micro-averaged metrics to minor class imbalances [41]. As for the temporal image classification, we predict one of "improving", "stable", and "worsening" for each one of the 5 findings: Consolidation, Pleural Effusion, Pneumonia, Pneumothorax and Edema. Following BioViL-T [5], we use macro-accuracy across the 5 classes to evaluate the performance.

4.2 Implementation Details

We set the minibatch size to 16 for single image-text pair training and to 4 for temporal training. Our model training was limited to a maximum of 5 studies per patient to accommodate resource limitations. We employed the AdamW [24] optimizer for model optimization. The learning rate was adjusted according to the training stage, with detailed strategies provided in the Supplementary Materials. The training was early stopped if the validation BLEU-4 score did not improve over 10 consecutive epochs. All experiments were conducted using two Nvidia GeForce RTX 3090 GPUs.

Table 1: Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics on MIMIC-CXR. The **Best** and second-best results of each metric are shown in **bold** and underline, respectively. † indicates the results are cited from their original papers. Since our study involves necessary data cleaning for longitudinal analysis, these results are not strictly comparable to ours. Results without † were obtained by re-running publicly available code on the same preprocessed dataset used in our study.

		NLG						CE		
Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F1
$\mathcal{M}^2$Transformer [11]	2019	0.352	0.211	0.138	0.096	0.128	0.263	0.239	0.173	0.173
R2Gen [10]	2020	0.339	0.211	0.143	0.103	0.138	0.279	0.297	0.189	0.193
R2GenCMN [9]	2021	0.345	0.213	0.143	0.101	0.140	0.274	0.354	0.271	0.275
$\mathcal{M}^2$TR.PROGRESSIVE [28]	2021	0.349	0.204	0.132	0.091	0.124	0.255	0.220	0.209	0.236
XProNet [48]	2022	0.303	0.188	0.127	0.091	0.128	0.268	**0.419**	0.230	0.242
CvT-212DistilGPT2 [27]	2022	<u>0.372</u>	<u>0.231</u>	<u>0.155</u>	<u>0.111</u>	<u>0.149</u>	<u>0.280</u>	<u>0.417</u>	<u>0.295</u>	<u>0.306</u>
DCL [21]	2023	0.263	0.153	0.099	0.071	0.117	0.211	0.303	0.232	0.229
HERGen(**Ours**)	2024	**0.395**	**0.248**	**0.169**	**0.122**	**0.156**	**0.285**	0.415	**0.301**	**0.317**
Results below are not strictly comparable due to our dataset preprocessing. For reference only.										
PPKED† [23]	2021	0.360	0.224	0.149	0.106	0.149	0.284	–	–	–
ContrastiveAttention† [25]	2021	0.350	0.219	0.152	0.109	0.151	0.283	0.352	0.298	0.303
AlignTransformer† [55]	2021	0.378	0.235	0.156	0.112	0.158	0.283	–	–	–
RGRG† [42]	2023	0.373	0.249	0.175	0.126	0.168	0.264	–	–	–
KIUT† [16]	2023	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
METransformer† [50]	2023	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311

4.3 Results of Radiology Report Generation

Results on MIMIC-CXR. Our model exhibits excellent radiology report generation capabilities, outperforming state-of-the-art models in both Natural Language Generation (NLG) and Clinical Efficiency (CE) metrics, as shown in Table 1. For NLG metrics, it notably surpasses all baseline models, notably improving over the second-best model, CvT-212DistilGPT2, by significant margins. Specifically, compared with CvT-212DistilGPT2, our model achieves a Δ+5.9% overall improvement on the averaged NLG metrics compared with CvT-212DistilGPT2. In CE metrics, our model enhances recall and F1 by $\Delta + 2.0\%$ and $\Delta + 3.6\%$, respectively, compared to the second-best results. Our precision score of 0.415 closely approaches the best score of 0.419. Additionally, we incorporate micro-based metrics for five common observations, following the methodologies of other studies [26, 42], to provide further evaluation of our method. These results are available in the Supplementary Material. Furthermore, our statistical analysis verifies that our model significantly outperforms the second-best approach, as detailed in Table 4.

Results on Longitudinal MIMIC-CXR. Table 2 presents a comparison of our model against various baseline methods in terms of Natural Language Generation (NLG) and Clinical Efficiency (CE) metrics. Our model outperforms both single-image and longitudinal-image-based methods in all evaluated NLG and CE metrics. Notably, our model achieves an increase of Δ+6.5% on the averaged NLG metrics compared with the second-best approach CvT-212DistilGPT2. In terms of CE metrics, our model also outperforms CvT-212DistilGPT2 in all

Table 2: Results of NLG metrics (BLEU (BL), METEOR (M), ROUGE-L (R_L)) and CE metrics (Precision, Recall and F1) on the Longitudinal MIMIC-CXR dataset. Results marked with a dagger (†) are cited from published literature Prefilling [59]. Since our curation of the Longitudinal-MIMIC dataset aligns with the approach in Prefilling [59], these results are directly comparable to ours.

	NLG						CE		
Method	BL-1	BL-2	BL-3	BL-4	M	R_L	Precision	Recall	F1
Baselines based on single images									
AoANet† [14]	0.272	0.168	0.112	0.080	0.115	0.249	-	-	-
CNN + Trans†	0.299	0.186	0.124	0.088	0.120	0.263	-	-	-
Transformer†	0.294	0.178	0.119	0.085	0.123	0.256	-	-	-
R2Gen† [10]	0.302	0.183	0.122	0.087	0.124	0.259	-	-	-
R2CMN† [9]	0.305	0.184	0.122	0.085	0.126	0.265	-	-	-
CvT-212DistilGPT2 [27]	0.365	0.226	0.151	0.107	0.143	0.275	0.367	0.258	0.261
Baselines based on longitudinal images									
Prefilling† [59]	0.343	0.210	0.140	0.099	0.137	0.271	-	-	-
HERGen(**Ours**)	**0.389**	**0.242**	**0.163**	**0.117**	**0.155**	**0.282**	**0.421**	**0.289**	**0.295**

cases, achieving the improvements of $\Delta + 14.7\%$ in precision, $\Delta + 12.0\%$ in recall, and Δ+13.0% in F1 score, respectively. Notably, our model also significantly surpasses longitudinal-image-based baseline [59], which also utilizes prior images and reports for current report generation, underscoring the effectiveness of our proposed temporal data integration strategy.

4.4 Results of Temporal Image Classification

The temporal image classification performance on MS-CXR-T is shown in Table. 3. We divided the dataset into training, validation, and test sets with a 70% / 10% / 20% ratio. In the finetuning phrase, we employ our pretrained image encoder and group causal transformer (these two modules remain frozen) to extract representations from pairs of images, and then only train a linear layer to make predictions. It is observed that HERGen achieves the best performance across 4 diseases and achieves the second-best performance on edema. Specifically, our model improve the macro-accuracy than the second best results by $\Delta + 11.1\%$, Δ+4.9%, Δ+14.7% and Δ+3.7% on consolidation, pleural effusion, pneumonia, pneumothorax, respectively. These advancements further underscore the effectiveness of our proposed group causal transformer in capturing the progression of diseases and extracting semantics from longitudinal studies.

4.5 Ablated Analysis of Our Framework

Effect of Auxiliary Contrastive Alignment. We delve into the impacts of incorporating our auxiliary contrastive alignment module, as delineated in Table 5. It is observed that incorporating the contrastive learning objective yields

Table 3: Temporal medical image classification performance on MS-CXR-T. Macro-accuracy ([%]) are used as the metric. The **Best** and second-best results are shown in **bold** and underline, respectively. Note that Pl.effusion denotes pleural effusion.

Method	Pre-train	Consolidation	Pl. effusion	Pneumonia	Pneumothorax	Edema
Random	-	32.3	31.6	30.3	39.0	34.9
ResNet	ImageNet	37.5	39.0	48.4	45.3	42.5
BioViL [7]	Static	42.9	41.4	47.9	42.8	40.7
BioViL-T [5]	Temporal	45.0	46.3	52.0	50.1	**52.0**
HERGen(**Ours**)	Temporal	**56.1**	**51.2**	**66.7**	**54.8**	48.1

Table 4: Comparison of CvT-212DistilGPT2 and HERGen with 95% confidence intervals, which are computed using non-parametric bootstrap.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
CvT-212DistilGPT2	0.372 (0.367, 0.377)	0.231 (0.227, 0.235)	0.155 (0.151, 0.158)	0.111 (0.107, 0.114)	0.280 (0.276, 0.283)	0.149 (0.147, 0.151)
HERGen(**Ours**)	0.396 (0.392, 0.400)	0.248 (0.244, 0.252)	0.168 (0.164, 0.172)	0.122 (0.118, 0.125)	0.285 (0.281, 0.288)	0.156 (0.154, 0.157)
Gains	**+0.023** (+0.017, +0.030)	**+0.017** (+0.011, +0.023)	**+0.014** (+0.008, +0.019)	**+0.011** (+0.005, +0.016)	**+0.005** (0.0, +0.01)	**+0.006** (+0.003, +0.009)

improvements in all NLG metrics compared to the baseline for both datasets, suggesting enhanced consistency in report generation. Notably, we observed a augmentation of +3.3% in average NLG metrics for the MIMIC-CXR dataset and a +2.3% improvement for the Longitudinal MIMIC-CXR dataset, underscoring the value of our contrastive alignment module in report generation.

Effect of Temporal Aggregation Module. We evaluate the impact of integrating our temporal aggregation module in Table 5. It is observed that this module significantly enhances the NLG metrics across the MIMIC-CXR and Longitudinal MIMIC-CXR datasets. Specifically, on the Longitudinal MIMIC-CXR dataset, it achieves +4.9% improvement in the averaged NLG metrics, compared to the baseline upon integrating this module. When combined with a model trained using contrastive learning, the improvement on the averaged NLG metrics compared with baseline further increases to +6.5%, which marks a significant enhancement than +2.3% (Row 2). This pattern is consistent across datasets, underscoring the temporal aggregation module's effectiveness in leveraging patient histories for generating more accurate reports.

Effect of Curriculum Learning Strategy. We evaluate the impact of our curriculum learning strategy on model performance in Table 5. Our analysis reveals that, across both the MIMIC-CXR and Longitudinal MIMIC-CXR datasets, the models incorporating contrastive learning alignment consistently outperform the baseline. Furthermore, our final model, which integrates both contrastive learning and temporal aggregation module, shows the best performance across the majority of metrics, highlighting the combined benefits of these approaches. For a detailed comparison between joint and curriculum-based training, please refer to the Supplementary Material.

Table 5: Ablation study of different components ("CL" represents the auxiliary contrastive alignment module, and "Temporal" denotes the group causal transformer for capturing longitudinal information). On each dataset, the row 1, 2 and 4 corresponds to the Stage 1, Stage 2, and Stage 3 of our curriculum learning strategy, respectively. Row 3 showcases a variant trained using our temporal approach without the contrastive learning component for comparison. The relative improvements in the average of all NLG metrics compared with baseline is presented in the "AVG.Δ" column.

Dataset	CL	Temporal	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	AVG.Δ
MIMIC-CXR			0.372	0.231	0.155	0.111	0.149	0.280	–
	✓		0.390	0.239	0.161	0.117	0.153	0.280	+3.3%
		✓	0.388	0.240	0.162	0.116	0.153	0.283	+3.4%
	✓	✓	**0.395**	**0.248**	**0.169**	**0.122**	**0.156**	**0.285**	+5.9%
Lon-MIMIC			0.365	0.225	0.151	0.107	0.143	0.275	–
	✓		0.375	0.232	0.155	0.110	0.146	0.277	+2.3%
		✓	0.380	0.237	0.160	0.115	0.153	**0.283**	+4.9%
	✓	✓	**0.389**	**0.242**	**0.163**	**0.117**	**0.155**	0.282	+6.5%

(a) CvT-212DistilGPT2 (b) Our approach

Fig. 5: Embedding visualization of MIMIC-CXR images in CvT-212DistilGPT2 and our model with t-SNE.

4.6 Qualitative Results

Case Study of Generated Reports. Fig. 6 presents a case study comparing reports generated by our model with those from CvT-212DistilGPT2 for a given patient. The comparison shows that reports from our model align more clinical findings with the ground truth. Moreover, our model correctly generates more comparative statements, such as "appear stable" or "appear unchanged", suggesting its superiority to capture temporal information. These findings underscore our model's proficiency in report generation by (1) identifying disease-specific features through consistent anatomical structures in patient-level CXRs and (2) generating time-comparative sentences.

Visualization of Learned Embeddings. Fig. 5 shows T-SNE visualization of image embeddings from a dataset constructed following 5×200 MIMIC-CXR dataset proposed in MedCLIP [51]. We calculated the 50-th percentile of the cosine similarity matrix within each class, revealing that our model achieves a higher average similarity (0.0919) compared to CvT-212DistilGPT2 (0.0245), indicating better disease-specific feature extraction.

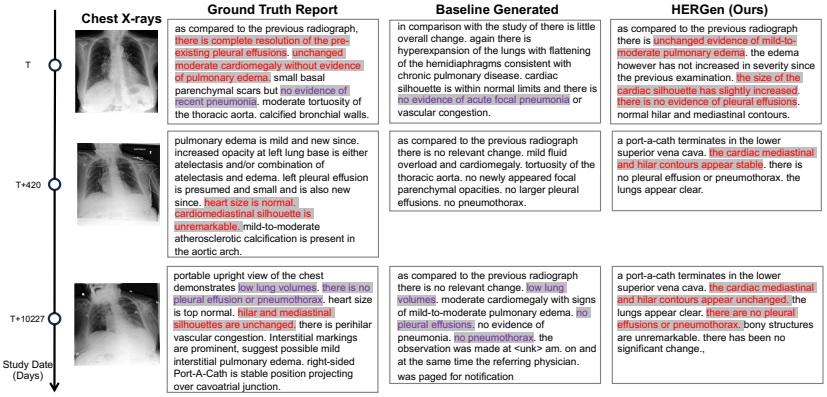

Fig. 6: This case study compares radiology report predictions for a patient by our model and CvT-212DistilGPT2. Text highlighted in gray indicates words or their synonyms found in both the predicted and ground truth reports. Purple highlights denote similar matches in the baseline-generated (CvT-212DistilGPT2) reports and ground truth, while red highlights show similar matches in our model's reports and ground truth. From top to bottom, the chest X-rays are chronologically ordered. Here T denotes the study date of the first study.

5 Conclusion

In this paper, we present a novel framework to enhance radiology report generation by utilizing the varying-size patient histories. By integrating a novel group casual transformer, our model effectively aggregates temporal information of longitudinal data. Besides, our framework optimize an auxiliary contrastive alignment module to further align image and textual data. Moreover, a curriculum learning strategy is employed to sequentially optimize these modules, thereby progressively improving model performance. Our extensive experiments demonstrate the model's capability to generate clinically precise reports and extract meaningful insights from historical data.

Limitations and Future Work. One potential limitation of our method is that the model's alignment operates within the embedding space without accounting for anatomical consistencies in longitudinal studies. Additionally, we plan to expand HERGen into a more comprehensive representation learning model, thereby broadening its utility across varied downstream tasks.

Acknowledgement

This work was partially supported by the Research Grants Council of Hong Kong (27206123 and T45-401/22-N), the Hong Kong Innovation and Technology Fund (ITS/273/22), and the National Natural Science Foundation of China (No. 62201483).

References

1. Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., Fahmy, A.: Automated radiology report generation using conditioned transformers. Informatics in Medicine Unlocked **24**, 100557 (2021)
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., de Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Schwaighofer, A., et al.: Ms-cxr-t: Learning to exploit temporal structure for biomedical vision-language processing (2023)
5. 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)
6. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
7. 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)
8. Cao, D.J., Hurrell, C., Patlas, M.N.: Current status of burnout in canadian radiology. Canadian Association of Radiologists Journal **74**(1), 37–43 (2023)
9. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
10. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
11. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
14. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4634–4643 (2019)
15. 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)
16. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19809–19818 (2023)

17. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. arXiv preprint arXiv:2004.12274 (2020)
18. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
19. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg-chest radiographs with structured labels. PhysioNet (2019)
20. 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)
21. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
22. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
23. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Ma, X., Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. arXiv preprint arXiv:2106.06965 (2021)
26. Miura, Y., Zhang, Y., Tsai, E.B., Langlotz, C.P., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042 (2020)
27. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
28. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. arXiv preprint arXiv:2102.09777 (2021)
29. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
31. Pavlopoulos, J., Kougia, V., Androutsopoulos, I., Papamichail, D.: Diagnostic captioning: a survey. Knowledge and Information Systems **64**(7), 1691–1722 (2022)
32. 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)
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
34. Ramesh, V., Chi, N.A., Rajpurkar, P.: Improving radiology report generation systems by removing hallucinated references to non-existent priors. In: Machine Learning for Health. pp. 456–473. PMLR (2022)
35. Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., Rosenow III, E.C.: Interpretation of plain chest roentgenogram. Chest **141**(2), 545–558 (2012)

36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
37. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online) **359** (2017)
38. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
39. Serra, F.D., Wang, C., Deligianni, F., Dalton, J., O'Neil, A.Q.: Controllable chest x-ray report generation from longitudinal representations. arXiv preprint arXiv:2310.05881 (2023)
40. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv:2004.09167 (2020)
41. Sorower, M.S.: A literature survey on algorithms for multi-label learning. Oregon State University, Corvallis **18**(1), 25 (2010)
42. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023)
43. Thrall, J.H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., Brink, J.: Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. Journal of the American College of Radiology **15**(3), 504–508 (2018)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
46. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
47. 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)
48. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. pp. 563–579. Springer (2022)
49. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9049–9058 (2018)
50. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023)
51. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
52. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22–31 (2021)

53. 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)
54. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
55. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021)
56. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4651–4659 (2016)
57. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)
58. 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)
59. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. arXiv preprint arXiv:2306.08749 (2023)