

# Divide and Conquer Radiology Report Generation via Observation Level Fine-grained Pretraining and Prompt Tuning

**Yuanpin Zhou**  
Zhejiang University  
HiThink Research  
zhouyuanpin@myhexin.com

**Huogen Wang**  
HiThink Research  
wanghuogen@myhexin.com

## Abstract

The automation of radiology report generation (RRG) holds immense potential to alleviate radiologists' workloads and improve diagnostic accuracy. Despite advancements in image captioning and vision-language pretraining, RRG remains challenging due to the lengthy and complex nature of radiology reports. In this work, we propose the Divide and Conquer Radiology Report Generation (DCRRG) model, which breaks down full-text radiology reports into concise observation descriptions. This approach enables the model to capture fine-grained representations from each observation through a two-stage process: an encoding stage focusing on observation prediction tasks to learn fine-grained representations, and a decoding stage for integrating these descriptions into cohesive and comprehensive radiology reports. Experimental results on two benchmark datasets demonstrate that DCRRG achieves significant improvements across all evaluation metrics, underscoring its capability to generate semantically coherent and clinically accurate radiology reports.

## 1 Introduction

Radiology images are commonly used to diagnose, monitor, and treat medical conditions in clinical practice (FDA, 2022). Recently, automatic radiology report generation (RRG) has garnered increasing attention from both the machine learning and medical fields. This technology aims to generate semantically coherent and informative reports to describe the corresponding examination images. Such techniques hold significant clinical potential by alleviating the workload of junior radiologists and reducing diagnostic errors through improved interpretation (Jing et al., 2018; Çallı et al., 2021).

Notable advancements in artificial intelligence have led researchers to propose various data-driven neural networks for automatic RRG, yielding promising results (Lu et al., 2017; Anderson

et al., 2018; Chen et al., 2020; Liu et al., 2021; Nooralahzadeh et al., 2021; Wu et al., 2022; Wang et al., 2022a; Li et al., 2023). Similar to the task of image captioning (Xu et al., 2015), which aims to describe the visual content of images, RRG models typically use an encoder-decoder architecture. In the encoding stage, visual representations of the images are extracted by a vision encoder, usually pretrained on image-label datasets. In the decoding stage, the medical report is generated by a decoder, employing either a Transformer (Chen et al., 2020; Liu et al., 2021) or LSTM (Jing et al., 2018) architecture with image-text datasets.

Recently, large-scale vision-language pretraining (VLP), such as CLIP (Radford et al., 2021), has achieved significant success through contrastive learning on image-text datasets. By jointly training on large-scale image-text pairs, this approach generates transferable representations that support versatile downstream tasks, enhancing the efficacy of vision encoders for RRG. Building on these advancements, recent studies (Zhang et al., 2020; Huang et al., 2021; Wang et al., 2022b, 2023) concentrate on refining vision encoders for RRG using contrastive learning with image-text paired datasets.

Despite the substantial advancements achieved in this field, RRG remains a difficult task. As is demonstrated in Figure 1, the radiology report comprise 6 sentences, describing 5 observations, among which 3 are positive, 1 is negative, and 1 is uncertain. The lengthy and complex nature of medical reports necessitates both the encoder and decoder be able to handle subtle and fine-grained representations. The accompanying bar chart displays the distribution of each observation, each facing varying degrees of data imbalance. However, since a report usually includes multiple observations, addressing the imbalance of one observation without impacting others presents a significant challenge.

In summary, several challenges significantly im-

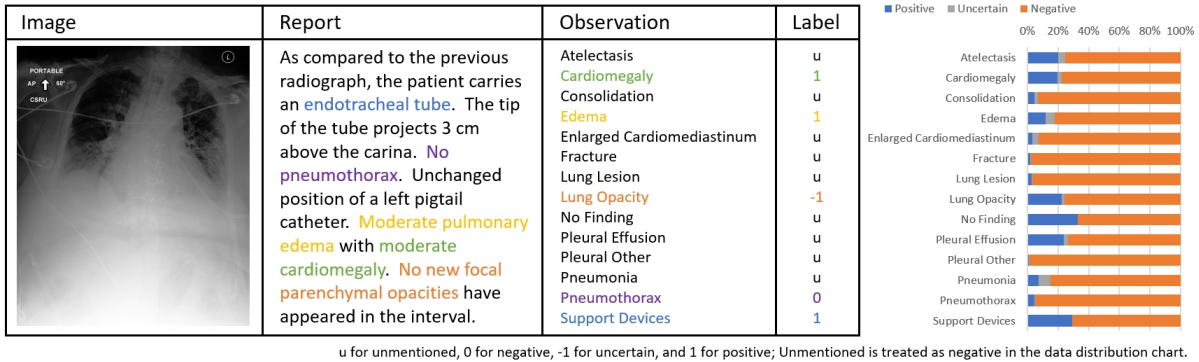


Figure 1: Demonstration of challenges in RRG. Left shows an example from the MIMIC-CXR dataset, including a X-ray image, a radiology report and the observations and corresponding labels extracted from the report. Right shows the data distribution of MIMIC-CXR based on 14 observation, respectively.

impact the effectiveness of RRG models: (1) Compared to image captioning, radiology reports are notably longer and comprise multiple sentences, each detailing specific medical observations. (2) Medical datasets frequently exhibit substantial class imbalances and skewed data distributions. (3) The scarcity of training data particularly limits the performance of data-intensive VLP models. Therefore, an ideal RRG model should: (1) Capably learn subtle and detailed representations both during the encoding and decoding stages. (2) Effectively address data imbalances for individual observations without unintended consequences for others. (3) Harness uni-modal data (such as image-only or text-only) alongside image-text pairs for enhanced training effectiveness.

Noticing that a fundamental distinction between RRG and image captioning lies in the constrained terminology used for describing diseases (observations) in radiology reports, compared to the open vocabulary typically encountered in image captioning tasks. Leveraging this distinction allows us to break down the lengthy and complex full-text radiology report into limited observation descriptions. The challenge then shifts to how to effectively learn subtle and nuanced representations from each observation and integrate them cohesively to generate the complete radiology report.

In response to this challenge, we propose the Divide and Conquer Radiology Report Generation (DCRRG) model. This approach focuses on two main stages: (1) Encoding Stage: The model learns fine-grained representations through observation prediction tasks. Each observation is treated independently to capture detailed nuances specific to that observation. (2) Decoding Stage: The model

generates a cohesive description for each indicated observation, integrating them to form the full-text radiology report.

Specifically, we employ prompts to bridge between individual observations and full reports. In the encoding stage, both the text and image encoders are trained with prompts to predict observation labels. This setup allows us to implement data balancing strategies and integrate uni-modal data effectively. Subsequently, the encoders undergo joint fine-tuning to align cross-modal representations via contrastive learning. In the decoding stage, the decoder is trained to generate descriptions for individual observations based on either image or text representations, guided by prompts indicating the specific observation to be generated. Prompt tuning further refines the model for efficient full-text report generation. In summary, our proposed DCRRG model contributes:

- We propose a novel method for RRG, employing a divide-and-conquer approach through observation-level training. This method optimizes the feature extraction network specifically for each individual observation, effectively mitigating the impact of data imbalance.
- To enhance the learning of uni-modal representations and facilitate integration between visual and textual modalities, we introduce divide-and-conquer contrastive learning (DCCLIP). This approach aims to refine accurate and fine-grained representations, thereby boosting performance in text-oriented medical report generation tasks.
- Experimental results on two benchmark datasets demonstrate that our proposed method achieves

significant improvements across all evaluation metrics.

## 2 Related Works

### 2.1 Medical Report Generation

Medical report generation involves interpreting medical images to generate comprehensive reports (Jing et al., 2018, 2019; Chen et al., 2020). Unlike image captioning (Xu et al., 2015; Anderson et al., 2018; Liu et al., 2018; Gu et al., 2022), which typically produces single-sentence descriptions for general images, medical report generation aims to generate paragraphs containing multiple clinical descriptions.

Inspired by the success of image captioning, several encoder-decoder-based frameworks have been introduced for medical report generation. For instance, (Jing et al., 2018) employed a hierarchical LSTM with an attention mechanism, while (Chen et al., 2020) utilized a Transformer to generate long paragraphs. (Yuan et al., 2019) explored methods to incorporate medical concepts to enhance performance. Additionally, (Jing et al., 2018) and (You et al., 2021) further integrated medical concepts into their models. (Yang et al., 2022), (Liu et al., 2021), and (Li et al., 2019) proposed approaches to construct medical knowledge graphs, injecting medical knowledge directly into the models.

In summary, deep learning models, especially encoder-decoder frameworks, have shown promising results in medical report generation, primarily trained end-to-end to produce full-text radiology reports. In our study, we adopt a different approach by decomposing the full-text RRG into tasks for generating descriptions of individual observations. We then fine-tune the model using prompt tuning techniques to generate comprehensive radiology reports.

### 2.2 Med-VLP

Vision-language pretraining (VLP) has demonstrated its capability to learn effective visual representations using image-caption pairs from general domains. Many approaches focus on learning visual-semantic embeddings for tasks such as vision-text retrieval (Liu et al., 2019; Wu et al., 2019; Lu et al., 2019; Huang et al., 2020; Chen et al., 2021) using attention or object detection models, as well as vision-text contrastive learning (Zhang et al., 2020; Jia et al., 2021; Yuan et al., 2021; Yu et al., 2022), and leveraging multiple forms of vision and text supervision (Singh

et al., 2021; Li et al., 2022). These methods are predominantly applied in general domains where vast amounts of web images and captions are available, far surpassing the scale of medical image-text datasets. This stark contrast poses challenges for applying self-supervised contrastive learning (CL) techniques to large-scale vision-text transformers in medical contexts. Although solutions such as data augmentation (Li et al., 2021) and knowledge graphs (Shen et al., 2022) have been proposed to mitigate these challenges, the disparity in data scale between general domains and medical domains remains substantial.

Inspired by this, medical vision-language pretraining (Med-VLP) was investigated based on contrastive learning as well (Zhang et al., 2020; Huang et al., 2021; Wang et al., 2021). Nonetheless, they all work on paired medical images and texts so still encounter the lacking data challenge. To resolve this, MedCLIP (Wang et al., 2022b) proposed to decouple images and texts for contrastive learning by recombining image-text pairs. Phenotype-CLIP (Wang et al., 2023) propose the phenotype-based contrastive learning to learn fine-grained representations.

Overall, the aforementioned Med-VLP models address data scarcity by augmenting training data through strategies like recombining or decomposing original paired datasets within a cross-modal training framework. In our study, we propose a method to initially learn uni-modal representations through pretraining. Subsequently, we fine-tune these representations under a contrastive learning framework to achieve cross-modal alignment. This approach allows for straightforward implementation of data balancing and complementation strategies during the pretraining phase.

## 3 Method

In this section, we will detail the implementation of our proposed Divide and Conquer Radiology Report Generation approach (DCRRG). The overall architecture of DCRRG is illustrated in Figure 2. Our model consists of four main steps. Step 1 and Step 2 are collectively referred to as DCCLIP, which constitutes a contrastive learning model for Med-VLP. Initially, DCCLIP pretrains on uni-modal data and subsequently finetunes on paired image-text datasets. Step 3 involves generating observation descriptions using uni-modal representations and prompts provided for each ob-

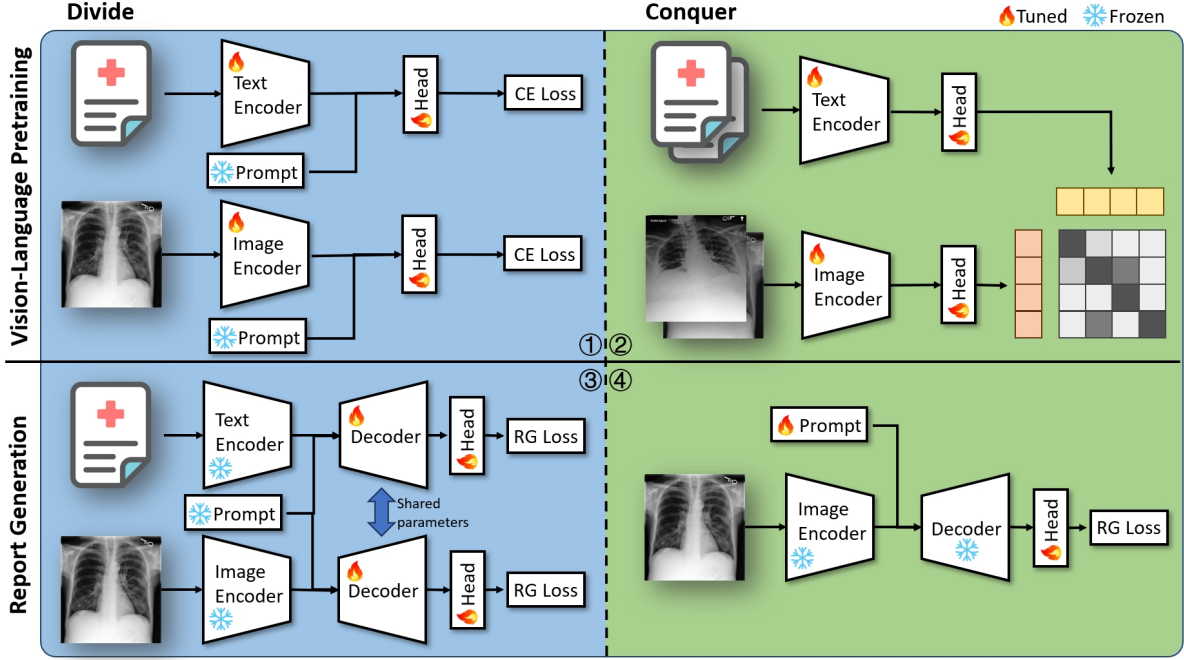


Figure 2: The overall architecture of DCRRG. (1) Observation level fine-grained pretraining. The encoders are trained based on single modality inputs to predict observation labels. (2) Contrastive Learning with calibrated semantic matching loss. (3) Single Observation description generation. (4) Full report generation with prompt tuning.

ervation. In Step 4, prompt tuning is applied to efficiently adjust the model parameters for full-text report generation. The details of each step is described in the following sections.

### 3.1 Observation Level Pretraining

In this step, we train the image encoder and text encoder using uni-modal data. The labels for uni-modal data are generated by an observation labeler, leveraging the CheXpert labeler (Irvin et al., 2019) in our study. For each image-text pair, a total of 14 observations are extracted. We treat unmentioned labels as negative and exclude data with uncertain labels from training. Optionally, we can apply data re-sampling for balancing and supplement the dataset with additional uni-modal data. To optimize the training process, we use Cross-Entropy Loss (CE Loss). This approach ensures effective learning and alignment of uni-modal representations during the initial training phase.

**Image Encoder.** We encode images into embeddings  $\mathbf{v} \in \mathbb{R}^D$  using a vision encoder  $E_{img}$ . A projection head then maps the raw embeddings and optional prompt tokens to  $\mathbf{v}_p \in \mathbb{R}^P$ .

$$\mathbf{v} = E_{img}(\mathbf{x}_{img}) \quad (1a)$$

$$\mathbf{v}_p = f_v(\text{concat}(\mathbf{v}, \mathbf{e}_{prompt})) \quad (1b)$$

where  $f_v$  is the projection head of the vision encoder.  $\mathbf{e}_{prompt} \in \mathbb{R}^{14}$  is the one-hot vector corresponding to 14 observations.

**Text Encoder.** Similarly, we create clinically meaningful text embeddings  $\mathbf{t} \in \mathbb{R}^M$  by a text encoder. We project them to  $\mathbf{t}_p \in \mathbb{R}^P$  as

$$\mathbf{t} = E_{txt}(\mathbf{x}_{txt}) \quad (2a)$$

$$\mathbf{t}_p = f_t(\text{concat}(\mathbf{t}, \mathbf{e}_{prompt})) \quad (2b)$$

where  $f_t$  is the projection head and  $E_{txt}$  denotes the text encoder. This gives the same embedding dimension  $P$  as the vision encoder.

### 3.2 Calibrated Semantic Matching Loss

In this step, we align cross-modal representations through contrastive learning using image-text pairs. Traditionally, InfoNCE loss in models like CLIP (Radford et al., 2021) has been effective but can lead to false negatives during training. To address this, approaches like Semantic Matching Loss (SM Loss), as proposed by Wang et al. (Wang et al., 2022b), create soft targets based on similarity between image and text labels. However, in medical contexts, observation descriptions often contain specific details about corresponding images. This specificity can challenge SM Loss’s ability to

distinguish samples with identical labels. To enhance its effectiveness, we propose calibrating SM Loss using uni-modal similarities. This adjustment helps SM Loss better capture nuanced distinctions present in medical image-text pairs, thereby improving alignment during contrastive learning.

During each iteration, with  $N_{batch}$  input images  $\{\mathbf{x}_{image}\}$  and text  $\{\mathbf{x}_{text}\}$  and the corresponding observation labels  $\mathbf{l}_{img}$  and  $\mathbf{l}_{txt}$ , the original soft targets  $s$  for SM Loss is defined by

$$s = \frac{\mathbf{l}_{img}^\top \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}. \quad (3)$$

For an image  $i$ , we obtain a set of  $s_{ij}$  where  $j = 1 \dots N_{batch}$  corresponds to the batch of texts. The calibrated soft target is computed by normalizing weighted  $s_{ij}$  across  $j$  by softmax.

$$y_{ij}^{v \rightarrow t} = \frac{\exp(w_{ij}s_{ij})}{\sum_{j=1}^{N_{batch}} \exp(w_{ij}s_{ij})}, \quad (4)$$

where  $w_{ij} = \text{sim}(\mathbf{t}_i, \mathbf{t}_j)/2 + 0.5$  and  $\text{sim}(\cdot)$  represents the cosine similarity. Similarly, the reversed text-to-image soft targets are obtained by

$$y_{ji}^{t \rightarrow v} = \frac{\exp(\hat{w}_{ji}s_{ji})}{\sum_{i=1}^{N_{batch}} \exp(\hat{w}_{ji}s_{ji})}. \quad (5)$$

where  $\hat{w}_{ji} = \text{sim}(\mathbf{v}_j, \mathbf{v}_i)/2 + 0.5$ . The logits are obtained by cosine similarities between image and text embeddings:

$$\hat{s}_{ij} = \tilde{\mathbf{v}}_i^\top \cdot \tilde{\mathbf{t}}_j, \quad (6)$$

where  $\tilde{\mathbf{v}}_i$  and  $\tilde{\mathbf{t}}_j$  are normalized  $\mathbf{v}_p$  and  $\mathbf{t}_p$ , respectively. The predicted similarity is also obtained by softmax function

$$\hat{y}_{ij} = \frac{\exp \hat{s}_{ij} / \tau}{\sum_{i=1}^{N_{batch}} \exp \hat{s}_{ij} / \tau}. \quad (7)$$

$\tau$  is the temperature initialized at 0.07. The *semantic matching loss* is hence the cross entropy between the logits and soft targets as

$$\mathcal{L}^{v \rightarrow l} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} y_{ij} \log \hat{y}_{ij}. \quad (8)$$

Likewise, we can compute  $\mathcal{L}^{l \rightarrow v}$  and then reach to

$$\mathcal{L}_{CSM} = \frac{\mathcal{L}^{v \rightarrow l} + \mathcal{L}^{l \rightarrow v}}{2} \quad (9)$$

as the final training objective.

### 3.3 Observation Description Generation

In this step, the decoder is trained to generate observation descriptions using given uni-modal representations and specific prompts. Due to alignment achieved in previous steps between text and image representations, text representations can be treated as an augmented view of image representations, effectively doubling the training data available to the model. Furthermore, since the full-text report generation task is decomposed into generating 14 observation descriptions, the total number of training samples is increased substantially—28 times more compared to conventional RRG models. This augmentation enhances the model’s ability to learn from diverse perspectives embedded within each observation, thereby improving the overall quality and coherence of generated radiology reports.

**Decoder.** Our report decoder consists of two Transformer decoder layers. The whole process of a decoder layer  $\mathbf{f}_d(\cdot)$  can be written as follows:

$$\mathbf{f}_d(\mathbf{y}) = \text{LN}(\text{FFN}(\mathbf{e}_{ca}) + \mathbf{e}_{ca}), \quad (10)$$

$$\mathbf{e}_{ca} = \text{LN}(\text{CA}(\mathbf{e}_{attn}, \mathbf{f}_I) + \mathbf{e}_{attn}), \quad (11)$$

$$\mathbf{e}_{attn} = \text{LN}(\text{MMHA}(\mathbf{y}) + \mathbf{y}), \quad (12)$$

where MMHA and CA represents the masked multi-head self-attention and cross attention mechanism in (Vaswani et al., 2017).  $\mathbf{y}$  is the input of decoder. In Cross-attention sublayer, for each head,  $\{\mathbf{Q}, \mathbf{K}^*, \mathbf{V}^*\}$  comes from  $\mathbf{Q} = W_q * \mathbf{e}_{attn}$ ,  $\mathbf{K} = W_k * \mathbf{f}_I$ , and  $\mathbf{V} = W_v * \mathbf{f}_I$ , where  $W_*$  are the learnable parameters. The  $\mathbf{f}_d(\mathbf{y})$  will be sent to a Linear & Log-Softmax layer to get the output of target sentences. Notably, only token embedding is adopted during the decoding procedure. The entire auto-regressive generation process can be written as follows:

$$p(T|I) = \prod_{t=1} p(y_t | y_1, \dots, y_{t-1}, I). \quad (13)$$

where  $y_t$  is the input token in time step  $t$ .

Typically, the report generation objective is the cross-entropy loss to compare the predicted token index sequence with the ground truth. Given the ground truth report  $\hat{T}$ , all the underlying modules are trained to maximize  $p(\mathbf{y}|I)$  by minimizing the following:

$$\mathcal{L}_{RG} = -\sum_{t=1}^{\hat{n}} \log p(\hat{y}_t | \hat{y}_1, \dots, \hat{y}_{t-1}, I). \quad (14)$$

### 3.4 Report Generation with Prompt Tuning

In the final step, we employ the prompt tuning technique, specifically P-Tuning v2 (Liu et al., 2022), to fine-tune the decoder for generating full-text radiology reports. Here, we freeze the parameters of the image encoder and the decoder that were pretrained in previous steps. Prompt tuning involves training a tunable soft prompt and a randomly-initialized classification head under the same objectives as in Step 3, where observation descriptions are generated from uni-modal representations and specific prompts. This process ensures that the decoder adapts efficiently to produce coherent and accurate full-text reports based on the learned representations.

## 4 Experiments

We assess our methods across two dimensions: the generated reports from DCRRG and the learned representations from DCCLIP. First, we outline two benchmark datasets, the metrics employed, and the evaluation settings. Subsequently, we present the primary findings and in-depth analysis of our approach on these datasets. The experimental settings can be found in Appendix A.

### 4.1 Datasets

**MIMIC-CXR** (Johnson et al., 2019) is a comprehensive chest X-ray database containing 377,110 images and 227,835 corresponding free-text radiology reports from the Beth Israel Deaconess Medical Center in Boston, MA. We follow (Chen et al., 2020) to pre-process the datasets: we adopt the official split to split the MIMIC-CXR dataset.

**CheXpert** (Irvin et al., 2019) is another extensive dataset comprising 224,316 chest X-rays collected from Stanford Hospital. Each X-ray is paired with its respective radiology report and labeled for the presence of 14 medical observations. The dataset encompasses data from 65,240 patients. For evaluation, we adopt the official dataset split designed for classification tasks. Additionally, following methodologies described in (Huang et al., 2021; Wang et al., 2022b), we sample a multi-class classification dataset named CheXpert-5x200 from the testing split. This subset contains 200 images exclusively labeled positive for the five CheXpert competition tasks: Atelectasis, Cardiomegaly, Edema, Pleural Effusion.

### 4.2 Baselines

**Conventional RRG Baselines.** To validate the effectiveness of our proposed method in RRG, we compare it against established image captioning approaches such as S&T (Vinyals et al., 2015), AdaAtt (Lu et al., 2017), and TopDown (Anderson et al., 2018), as well as those specifically designed for the medical domain including R2Gen (Chen et al., 2020), PPKED (Liu et al., 2021), M2TR (Nooralahzadeh et al., 2021), DeltaNet (Wu et al., 2022), XProNet (Wang et al., 2022a), and DCL (Li et al., 2023). Results from these baselines are cited directly from their respective original papers, ensuring consistency with the reported settings.

**VLP Baselines.** We extend our evaluation to include a diverse set of VLP models: CONVIRT (Zhang et al., 2020), GLoRIA (Huang et al., 2021), MedCLIP (Wang et al., 2022b) and PhenotypeCLIP (Wang et al., 2023). For the decoding stage in downstream RRG tasks, we implement settings consistent with those described in (Wang et al., 2023). Additionally, we include CLIP (Radford et al., 2021) for evaluating DCCLIP, omitting PhenotypeCLIP (Wang et al., 2023) due to lack of public codebase and reported results in image-text retrieval and classification tasks.

### 4.3 Evaluation Metrics

For evaluating report generation, we utilized standard metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), computed with a standard evaluation toolkit. While BLEU and METEOR are commonly used in machine translation, ROUGE-L assesses summary quality. To assess disease prediction accuracy in report generation, we employed clinical efficacy (CE) metrics. These metrics, such as precision, recall, and F1 scores, were computed using disease labels extracted from both real reports and model predictions from the CheXpert dataset (Irvin et al., 2019).

For evaluating DCCLIP, we employed specific metrics tailored to its tasks. We utilized the Precision@K metric in the image-text retrieval task, which measures precision within the top K retrieved reports. This assessment determines if the selected report matches the category of the query image in the image-text retrieval process. Additionally, for evaluating downstream classification tasks,

Table 1: Experimental results of our model and baselines on the MIMIC-CXR dataset. The best results are in boldface. The VLP baselines are highlighted in platinum. DCRRG w/o balancing denotes the model is trained without data balancing technique. DCRRG w/ CheXpert denotes the model is trained with data complementation from CheXpert. \* denotes the re-implementations of existing contrastive learning methods for medical report generation.

Model	NLG Metrics				CE Metrics		
	B-4	MTR	R-L	CIDEr	P	R	F <sub>1</sub>
S&T (Vinyals et al., 2015)	0.084	0.124	0.263	-	0.249	0.203	0.204
AdaAtt (Lu et al., 2017)	0.088	0.118	0.266	0.084	0.268	0.186	0.181
TopDown (Anderson et al., 2018)	0.092	0.128	0.267	0.073	0.320	0.231	0.238
R2Gen (Chen et al., 2020)	0.103	0.142	0.277	0.253	0.333	0.273	0.276
PPKED (Liu et al., 2021)	0.106	0.149	0.284	0.237	-	-	-
M2TR (Nooralahzadeh et al., 2021)	0.107	0.145	0.272	-	0.240	0.428	0.308
DeltaNet (Wu et al., 2022)	0.114	-	0.277	0.281	-	-	-
XProNet (Wang et al., 2022a)	0.105	0.138	0.279	-	-	-	-
DCL (Li et al., 2023)	0.109	0.150	0.284	0.281	0.471	0.352	0.373
ConVIRT* (Zhang et al., 2020)	0.104	0.142	0.279	0.255	0.333	0.275	0.281
GLoRIA* (Huang et al., 2021)	0.106	0.143	0.281	0.253	0.347	0.276	0.306
MedCLIP* (Wang et al., 2022b)	0.109	0.146	0.283	0.255	0.353	0.312	0.322
PhenotypeCLIP (Wang et al., 2023)	0.119	0.158	0.286	0.259	-	-	-
DCRRG	0.119	0.161	0.284	0.282	0.460	0.314	0.376
DCRRG w/o balancing	0.116	0.156	0.283	0.276	0.413	0.312	0.354
DCRRG w/ CheXpert	<b>0.120</b>	<b>0.163</b>	<b>0.291</b>	<b>0.283</b>	<b>0.489</b>	<b>0.341</b>	<b>0.401</b>

we used the Area Under the ROC Curve (AUROC). This metric assesses the model’s ability to distinguish between different classes based on its ROC curve performance.

## 4.4 Experimental Results

### 4.4.1 Radiology Report Generation

The results on the MIMIC-CXR datasets are presented in Table 1, where we compare our proposal with existing state-of-the-art RRG models. Additionally, we implemented three VLP methods, namely ConVIRT (Zhang et al., 2020), MedCLIP (Wang et al., 2022b), and GLoRIA (Huang et al., 2021). The competitive performance of these VLP models alongside dedicated medical report generation models underscores the importance of learning fine-grained representations for generating lengthy and complex reports.

**Descriptive Accuracy** The proposed DCRRG method demonstrated promising performance on the MIMIC-CXR benchmark datasets. R2Gen (Chen et al., 2020) has recently been widely used as a baseline for MRG models. PPKED (Liu et al., 2021) integrates medical knowledge with typical MRG frameworks. Other baseline models, such as M2TR (Nooralahzadeh et al., 2021) and TopDown (Anderson et al., 2018), are also included

for comparison. As shown in Table 1, our DCRRG achieves state-of-the-art descriptive accuracy, outperforming others in METEOR and CIDEr metrics while matching their performance in BLEU-4 and ROUGE-L metrics. Higher CIDEr values indicate that our model generates reports with more coherent content topics, avoiding repetitive sentences seen in the training set. When trained with additional data from CheXpert for complementation, our model consistently surpasses all previous RRG methods across all evaluation metrics.

**Clinical Efficacy** We also evaluate our method by clinical efficacy (CE) metrics on the MIMIC-CXR dataset and compare the performances with other baseline models. Following official splitting, we directly cite the results from original papers for comparison. The experimental results in Table 1 reveal that our DCRRG significantly outperforms the previous models on three CE metrics. The experimental results in Table 1 demonstrate that DCRRG significantly outperforms previous models on three CE metrics. Compared to current state-of-the-art methods that leverage general and specific knowledge, our approach shows a notable performance improvement. This enhancement underscores the importance of our method and confirms its ability to predict more accurate clinical information.

Table 2: Results of Image-Text retrieval tasks on CheXpert-5x200 dataset. We take the Precision@{1,2,5,10} to measure the performance of various models in this task. Best within the data are in bold. DCCLIP w/o ba denotes the model is trained without data balancing technique.

Model	P@1	P@2	P@5	P@10
CLIP	0.20	0.20	0.21	0.20
ConVIRT	0.21	0.20	0.20	0.19
GLoRIA	0.46	0.48	0.47	0.46
MedCLIP	0.45	0.50	0.48	0.49
DCCLIP	<b>0.48</b>	<b>0.50</b>	<b>0.51</b>	<b>0.51</b>
DCCLIP w/o ba	0.45	0.49	0.49	0.49

#### 4.4.2 Image-text Retrieval

Following (Zhang et al., 2020; Huang et al., 2021; Wang et al., 2022b), we use the CheXpert-5x200 dataset to evaluate the effectiveness of our representation learning framework for image-text retrieval. Given an image as the input query, we retrieve target reports by computing the similarity between the query image and all candidate reports using the learned representations. Precision@K metric is employed to measure the precision in the top K retrieved reports, ensuring that the selected report belongs to the same category as the query image.

Based on the results presented in Table 2, it is evident that our method achieves superior performance compared to all other methods. This underscores the efficiency of our approach in providing necessary semantic information for text retrieval. Moreover, DCCLIP achieves competitive performance even without employing data balancing techniques, highlighting the benefits of observation-level pretraining compared to state-of-the-art (SOTA) methods.

#### 4.4.3 Classification

Following (Zhang et al., 2020), we assess each pre-trained image encoder under two distinct settings: a linear classification setup, where the pretrained CNN weights remain frozen and only a linear classification head is trained; and a fine-tuning scenario, where both the CNN weights and the linear head undergo fine-tuning. These settings complement each other in evaluation: the linear setup directly gauges the quality of the extracted image features using the pretrained CNN, while the fine-tuning setup mirrors how the pretrained CNN weights are typically utilized in practical applications. To further

Table 3: Results of classification (AUROC score) on CheXpert test sets based on different portion of training data: 1%, 10%, 100%. Best within the data are in bold. DCCLIP w/o ba denotes the model is trained without data balancing technique.

Model	Linear			Finetuning		
	1%	10%	100%	1%	10%	100%
Random	56.1	62.6	65.7	70.4	80.7	85.4
ImageNet	74.4	79.1	81.4	80.1	84.3	87.1
CLIP	80.2	82.3	83.1	82.2	84.8	87.6
ConVIRT	85.9	86.8	87.3	87.0	88.1	88.3
GLoRIA	86.6	87.8	88.1	87.0	88.1	88.3
MedCLIP	86.8	87.6	88.3	87.1	87.9	88.2
DCCLIP	<b>87.2</b>	<b>88.0</b>	<b>88.4</b>	<b>88.8</b>	<b>89.3</b>	<b>89.3</b>
DCCLIP w/o ba	86.8	87.7	88.2	88.1	89.0	89.1

compare the data efficiency of various pretraining methods, we evaluate the image encoders with 1%, 10%, and 100% of the training data, respectively, in each setting.

Table 3 presents the classification results for CheXpert across varying percentages of training data. To mitigate the impact of result variance due to random sampling of training data, we averaged the outcomes over five independent runs. We use the area under the ROC curve (AUROC) as our evaluation metric. Our method achieves superior performance compared to SOTA methods in both settings. Notably, even when trained with only 1% of the data, our method consistently outperforms models initialized with ImageNet weights trained on 100% of the data in both settings. This underscores the efficacy of DCCLIP in learning fine-grained representations for label-efficient classification.

#### 4.5 Ablation Study

We performed ablation experiments to demonstrate the effectiveness and various impacts of the intermediate steps in our proposed method. The experimental results are shown in Table 4. In summary, we found that both Step 1 and Step 2 are essential for descriptive accuracy and clinical efficacy in radiology report generation (RRG), with Step 1 having a significant impact on clinical efficacy. Step 3 is important for descriptive accuracy but contributes little to clinical efficacy. Utilizing Prompt Tuning (PT) is competitive with Full-parameters Fine-Tuning (FFT) in Step 4. The detailed key findings of the ablation study are listed in Appendix B.




Image	Ground Truth	PhenotypeCLIP	DCRRG
	<p><b>Bilateral pleural catheters</b> remain in place, with persistent pneumothoraces, moderate left apical lateral <b>pneumothorax</b> on the left and small on the right. The left <b>pneumothorax</b> is unchanged, but right <b>pneumothorax</b> has minimally increased. <b>Heart size remains normal</b>. Persistent left basilar <b>atelectasis</b> and adjacent small left <b>pleural effusion</b>.</p>	<p>Lung volumes remain low. The appearance of the cardiac silhouette is unchanged. There is a moderate <b>pneumothorax</b> evidenced in the right. Small bilateral <b>pleural effusions</b> are present. The mediastinal contours are within normal limits.</p>	<p>There is little change in the <b>support devices</b>. Moderate <b>pneumothorax</b> is found on the left and increased on the right. <b>The cardiomeastinal silhouette is normal</b>. There are no findings of focal consolidation, edema or pneumonia. There is mild left base <b>atelectasis</b> and bilateral <b>pleural effusions</b> seen on the frontal view.</p>

Figure 3: Qualitative comparison with a strong baseline model PhenotypeCLIP. We adopt the bold text to denote the observation key words in the radiology reports.

#### 4.6 Pipeline Error Analysis

The ablation study demonstrates the importance of Step 1 in DCRRG. To evaluate the potential impact of errors in observation labeling, we conducted a pipeline error analysis to assess the correlation between classification accuracy and final report generation performance. We used the four variant models in Step 1 (i.e., 6a 6d) to simulate various degrees of pipeline errors. We measured the Area Under the ROC Curve (AUROC) for predicting each observation on the MIMIC-CXR testing set, and also calculated the average AUROC for each model. More details can be found in Appendix C.

#### 4.7 Qualitative Analysis and Case Study

We further conducted a qualitative analysis to give a better understanding of our method. Specifically, we show a medical report generated by a strong baseline model PhenotypeCLIP and our proposed DCRRG. As is shown in Figure 3, our method can generate reports with better clinical efficacy than PhenotypeCLIP. Although PhenotypeCLIP correctly describes the observation of "cardiomegaly", "pneumothorax" and "pleural effusions", it failed to capture the observation of "support device" and "basilar atelectasis". Our proposed DCRRG is able to correctly describe all five observation in this example. One limitation shown in this example is that DCRRG failed to integrally described the observation of "pneumothorax", since our method tends to generate reports with one sentence for each observation, while the ground truth report has multiple sentences describing the observation of "pneumothorax".

We further utilize this example for case study and demonstrate the intermediate results of our

proposed DCRRG. As is shown in Table 7, the image encoder and the corresponding classification head trained in Step 1 is able to accurately predict the observation label. The corresponding single observation descriptions generated in Step 3 are also demonstrated. Each observation is described in one sentence and the observation descriptions are consistent with the observation labels. As we further compare the observation description with the generated report, we notice that there could be redundant sentences in the generated report. For example, "There are no findings of focal consolidation, edema or pneumonia." describe the unmentioned observations that did not occur in the ground truth report. We believe that the redundancy in the generated reports could potentially improve clinical efficacy but might degrade descriptive accuracy of DCRRG.

## 5 Conclusion

In this study, we introduce DCRRG, a divide and conquer training strategy for radiology report generation. This approach optimizes the feature extraction network individually for each observation, effectively mitigating data imbalance. To enhance the learning of uni-modal representations and facilitate integration between visual and textual modalities, we propose divide-and-conquer contrastive learning (DCCLIP). This method aims to refine accurate and fine-grained representations, thereby improving performance in text-oriented medical report generation tasks. Experimental results on two benchmark datasets show that our approach achieves substantial enhancements across all evaluation metrics.

## Limitations

Our framework has several limitations. Firstly, since observations guide our model, accurate labeling by observation extraction tools is crucial. Inaccuracies in this process may introduce biases. Future work will explore leveraging Large Language Models (LLMs) for more reliable observation extraction. Secondly, our framework operates as a pipeline, where the performance of report generation heavily depends on VLP accuracy. Consequently, errors can accumulate, especially with smaller datasets. Lastly, our framework specializes in generating radiology reports from Chest X-ray images. Future investigations should extend its applicability to other medical image types.

## Ethics Statement

The MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019) datasets used in our study are publicly available, ensuring no protected health information is disclosed. However, any inaccuracies in the generated reports, such as misdiagnoses or missed abnormalities, can lead to incorrect clinical outcomes. Therefore, it is crucial to control the use of model-generated reports and ensure that medical professionals review and validate them in clinical practice. Similar to other deep learning models, DCCLIP is susceptible to inherent biases present in the training data. It is essential to address fairness concerns and mitigate potential biases. Therefore, we strongly recommend users to carefully consider the ethical implications of the generated outputs in real-world applications.

## References

- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. [Publicly available clinical bert embeddings](#). *arXiv preprint*.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.
- Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.
- Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. 2021. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125.
- Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. [Generating radiology reports via memory-driven transformer](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- FDA. 2022. [Medical imaging](#).
- Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. [PPT: Pre-trained prompt tuning for few-shot learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#). *arXiv preprint*.
- Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.
- Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. [Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers](#). *arXiv preprint arXiv:2004.00849*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

- Baoyu Jing, Zeya Wang, and Eric Xing. 2019. [Show, describe and conclude: On exploiting the structure information of chest X-ray reports](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy. Association for Computational Linguistics.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. [On the automatic generation of medical imaging reports](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.
- Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6666–6673.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.
- Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.
- Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Fenglin Liu, Yuanxin Liu, Xuancheng Ren, Xiaodong He, and Xu Sun. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. *Advances in Neural Information Processing Systems*, 32.
- Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018. [simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 137–149, Brussels, Belgium. Association for Computational Linguistics.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. [P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. [Progressive transformer-based generation of radiology reports](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. 2022. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. FLAVA: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

- you need. *Advances in neural information processing systems*, 30.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.
- Jun Wang, Abhir Bhalerao, and Yulan He. 2022a. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer.
- Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. 2023. [Fine-grained medical vision-language representation learning for radiology report generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15949–15956, Singapore. Association for Computational Linguistics.
- Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. 2021. Self-supervised image-text pre-training with mixed data in chest x-rays. *arXiv preprint arXiv:2103.16022*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. [MedCLIP: Contrastive learning from unpaired medical images and text](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618.
- Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S. Kevin Zhou, and Li Xiao. 2022. [DeltaNet: Conditional medical report generation for COVID-19 diagnosis](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2952–2961, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. [Show, attend and tell: Neural image caption generation with visual attention](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.
- Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.
- Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. 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*, pages 72–82. Springer.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.
- Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer.
- Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

## A Experimental Settings

We utilized ResNet-50 (He et al., 2015), pretrained on ImageNet (Deng et al., 2009), as the image encoder to extract visual features from input medical images, following the approach in (Zhang et al., 2020). For textual features of input reports and sentences, we employed BERT initialized with ClinicalBERT weights (Alsentzer et al., 2019). The BERT model was configured with 8 attention heads and a hidden size of 512. During the training of DCCLIP in the encoding stage, we employed the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 1e-4, weight decay of 1e-6, and a batch size of 32. For training the decoder in Steps 3 and 4, we also used the AdamW optimizer with a learning rate of 1e-4 and a batch size of 16. During RRG inference, we employed a beam search of size 3 to improve the quality of generated outputs.

Table 4: Ablation study on the intermediate steps of DCRRG. Both NLG Metrics and CE Metrics are used to evaluate the final generated text reports. Notice that Prompt Tuning (PT) in Step 4 must be bundled together with Step 3, hence Full-parameters Fine-Tuning (FFT) is adopted when Step 3 is removed from the pipeline of DCRRG, where both the image encoder and the decoder are tunable. All models are trained and tested solely on the MIMIC-CXR datasets.

Model No.	Step 1	Step 2	Step 3	Step 4	B-4	MTR	R-L	CIDEr	P	$F_1$
1	✓	×	×	FFT	0.112	0.148	0.282	0.257	0.436	0.363
2	×	✓	×	FFT	0.110	0.145	0.283	0.257	0.357	0.324
3	×	✓	✓	PT	0.113	0.150	0.280	0.276	0.360	0.327
4	✓	×	✓	PT	0.116	0.156	0.282	0.279	0.438	0.365
5	✓	✓	×	FFT	0.115	0.156	0.281	0.274	0.456	0.371
6	✓	✓	✓	PT	0.119	0.161	0.284	0.282	0.460	0.376
7	✓	✓	✓	FFT	0.120	0.160	0.284	0.283	0.459	0.376

Table 5: Ablation study on the important techniques used in Step 1, where balancing refers to re-sampling data to balance labels for each observation and extra data refers to using additional data from the CheXpert dataset for training. We keep Step 2, Step 3 and Step 4 under the default settings and evaluate the final generated text reports with both NLG Metrics and CE Metrics.

Model No.	balancing	extra data	B-4	MTR	R-L	CIDEr	P	R	$F_1$
6a	×	×	0.116	0.156	0.283	0.276	0.413	0.312	0.354
6b	✓	×	0.119	0.161	0.284	0.282	0.460	0.314	0.376
6c	×	✓	0.117	0.158	0.283	0.280	0.446	0.313	0.368
6d	✓	✓	0.120	0.163	0.291	0.283	0.489	0.341	0.401

Table 6: Area Under the ROC Curve (AUROC) for the prediction of each observation of different models.

Observation	6a	6b	6c	6d
Atelectasis	0.62	0.63	0.64	0.65
Cardiomegaly	0.63	0.64	0.64	0.65
Consolidation	0.65	0.74	0.67	0.78
Edema	0.76	0.77	0.77	0.79
Enlarged Cardiomedastinum	0.61	0.70	0.63	0.76
Fracture	0.62	0.72	0.64	0.78
Lung Lesion	0.65	0.75	0.65	0.81
Lung Opacity	0.78	0.79	0.78	0.80
No Finding	0.74	0.74	0.75	0.74
Pleural Effusion	0.81	0.82	0.83	0.84
Pleural Other	0.62	0.79	0.62	0.85
Pneumonia	0.72	0.76	0.73	0.79
Pneumothorax	0.68	0.79	0.69	0.81
Support Devices	0.80	0.82	0.83	0.83
Average AUROC	0.69	0.75	0.71	0.78

## B Ablation Study

The key findings of the ablation study are listed below:

- Comparing Model 3 and Model 6 shows significant degradation in both NLG and CE Metrics when Step 1 is removed. Additionally, comparing Model 2 and Model 5 highlights the importance of Step 1 for performance.

- Comparing Model 4 and Model 6 reveals that Step 2 improves both NLG and CE Metrics. However, comparing Model 1 and Model 5 indicates that Step 2 is particularly important for NLG Metrics when Step 3 is removed.
- Comparing Model 5 and Model 6 shows that Step 3 is essential for improving NLG Metrics but not as important for CE Metrics. A similar observation is made by comparing Model 2 and Model 3.
- Comparing Model 6 and Model 7 demonstrates that PT matches the performance of FFT, with PT using only 0.1% to 3% of trainable parameters per task compared to FFT. This substantially reduces training time, memory cost, and per-task storage cost.

We further examined how various techniques in Step 1 impact the performance of DCRRG. As shown in Table 5, Model 6d achieves the best performance when both balancing and additional data are applied during training. Additionally, we found that balancing is more important than adding extra data for training.

Table 7: Intermediate results of our proposed DCRRG for the example of study s52926904. Here, label is denoted as the label extracted by CheXpert from the ground true report. Probability is denoted as the probability for each observation predicted by the image encoder and the corresponding classification head in Step 1. Observation description is denoted as the single observation description generated in Step 3.

Observation	Label	Probability	Observation description
Atelectasis	1	0.76	There is mild left base atelectasis seen on the frontal view.
Cardiomegaly	0	0.23	The cardiomediastinal silhouette is normal.
Consolidation	u	0.19	There is no focal consolidation.
Edema	u	0.27	The lungs are clear of edema.
Enlarged Cardiomeastinum	u	0.30	The cardiomeastinum is unremarkable.
Fracture	u	0.12	There is no visualized displaced rib fracture.
Lung Lesion	u	0.67	Numerous bilateral lesions are found on the frontal view.
Lung Opacity	u	0.12	There is no focal opacity to suggest pneumonia.
No Finding	u	0.02	There is no finding for this study.
Pleural Effusion	1	0.81	There is pleural effusions.
Pleural Other	u	0.10	There is no finding of pleural other.
Pneumonia	u	0.10	No signs of pneumonia.
Pneumothorax	1	0.73	Moderate right lateral pneumothorax is new.
Support Devices	1	0.88	There is little change in the monitoring and support devices.

### C Pipeline Error Analysis

As shown in Table 6, Model 6d achieved the highest average AUROC of 0.78 across 14 observations, while Model 6a achieved the lowest average AUROC of 0.69. By comparing the AUROC in Table 6 with the NLG and CE Metrics in Table 5, we observed a positive correlation between classification accuracy and final report generation performance. However, DCRRG does not require classification performance to be extremely accurate. Model 6b outperformed state-of-the-art methods with an average AUROC of 0.75. Additionally, examining the AUROC for each observation prediction task, we noticed significant improvements in extremely imbalanced observations when using balancing techniques (Model 6b and Model 6d). We also found that training with extra data (Model 6c) is not as helpful as balancing (Model 6b).