



# Act Like a Radiologist: Radiology Report Generation across Anatomical Regions

Qi Chen<sup>1\*</sup>, Yutong Xie<sup>1\*</sup>, Biao Wu<sup>2</sup>, Xiaomin Chen<sup>3</sup>, James Ang<sup>4</sup>  
Minh-Son To<sup>1,4,5</sup>, Xiaojun Chang<sup>2</sup>, and Qi Wu<sup>1†</sup>

<sup>1</sup> Australian Institute for Machine Learning, University of Adelaide  
<sup>2</sup> University of Technology Sydney, <sup>3</sup> South China University of Technology  
<sup>4</sup> Royal Adelaide Hospital, <sup>5</sup> Flinders University

**Abstract.** Automating radiology report generation can ease the reporting workload for radiologists. However, existing works focus mainly on the chest area due to the limited availability of public datasets for other regions. Besides, they often rely on naive data-driven approaches, *e.g.*, a basic encoder-decoder framework with captioning loss, which limits their ability to recognise complex patterns across diverse anatomical regions. To address these issues, we propose X-RGen, a radiologist-minded report generation framework across six anatomical regions. In X-RGen, we seek to mimic the behaviour of human radiologists, breaking them down into four principal phases: 1) initial observation, 2) cross-region analysis, 3) medical interpretation, and 4) report formation. Firstly, we adopt an image encoder for feature extraction, akin to a radiologist’s preliminary review. Secondly, we enhance the recognition capacity of the image encoder by analysing images and reports across various regions, mimicking how radiologists gain their experience and improve their professional ability from past cases. Thirdly, just as radiologists apply their expertise to interpret radiology images, we introduce radiological knowledge of multiple anatomical regions to further analyse the features from a clinical perspective. Lastly, we generate reports based on the medical-aware features using a typical auto-regressive text decoder. Both natural language generation (NLG) and clinical efficacy metrics show the effectiveness of X-RGen on six X-ray datasets. Our code and checkpoints are available at: <https://github.com/YtongXie/X-RGen>.

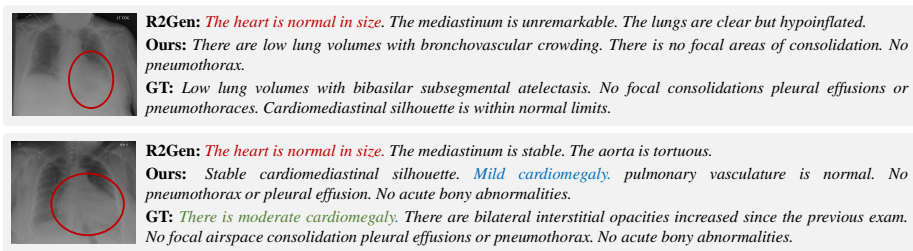
**Keywords:** Radiology Report Generation · Multiple Anatomical Regions · Radiologist-minded Framework

## 1 Introduction

The tasks of interpreting radiology images and producing reports are both arduous and prone to errors. To reduce this burden, automatic report generation systems can provide candidate reports for radiologists to verify. Besides, these systems can leverage data-hungry machine learning paradigms by learning directly

---

\* Equal contributions. † Corresponding author.



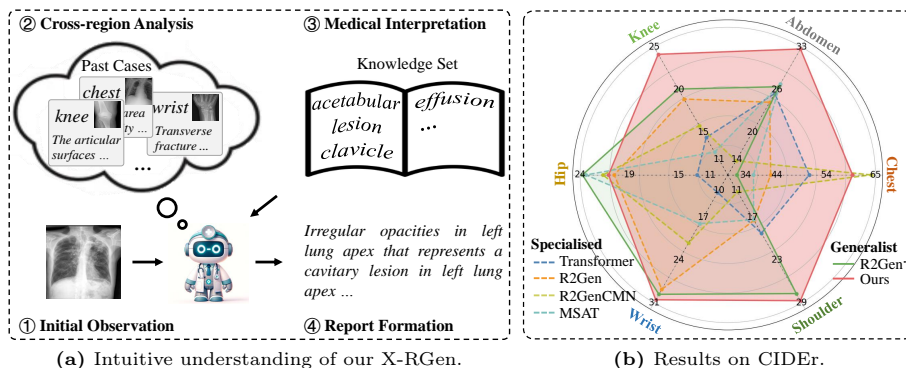
**Fig. 1:** Reports written by radiologists vs. existing models (e.g., R2Gen [5] trained on our merged dataset), and our X-RGen. We observe that R2Gen remembers some commonly used descriptions (highlighted in red) regardless of the semantic alignment with images, e.g., the correct diagnosis is “there is moderate cardiomegaly” (highlighted in green) while R2Gen keeps “the heart is normal in size” (highlighted in red).

from free-text reports, which is a significant advantage compared to other medical image analysis applications (e.g., medical image segmentation [15, 20, 36, 48]) that often rely on large amounts of quality annotations.

Radiologists commonly write reports based on radiology images covering different body parts. Despite notable progress, existing report generation works [22, 24, 25, 28, 38, 45, 49] have primarily focused on the chest, a limitation stemming from the scarcity of publicly available datasets for other anatomical regions. This narrow focus hampers the broader clinical utility of these systems. Besides, as they are designed following the typical single-dataset training-and-testing paradigm, they inevitably suffer from severe performance drop issues, when these generation models are directly deployed to another dataset w.r.t. various body regions. By contrast, learning across various anatomical regions can potentially uncover underlying commonalities in medical images, e.g., the fracture in the wrist, shoulder, knee and other parts; or overlapping areas in chest and abdomen X-ray images. Thus, it is crucial to design a report generation framework capable of covering multiple anatomical regions.

Technically, the heavy reliance on naive data-driven methods, such as basic encoder-decoder frameworks only with simple captioning loss, curtails their capability to identify complex medical patterns. In this way, not all the generated reports are semantically consistent with the images as the model tends to remember an “average” version that contains the frequently occurring words and phrases present in the training corpus [3] (see Figure 1). However, in radiology reports, there are many rare but critically important medical terminology vital for diagnosis. Thus, the challenge lies in designing a model that not only captures the typical data patterns but also recognises and accurately incorporates critical, albeit infrequent, medical terminology into radiology reports, ensuring a high degree of semantic consistency and diagnostic relevance.

To address these, we propose a radiologist-minded framework for generating radiology reports across diverse anatomical regions, named X-RGen. It covers six body parts: chest, shoulder, hip, knee, abdomen, and wrist. From Figure 2a, our



**Fig. 2:** (a) X-RGen mimics the behaviour of how human radiologists write reports. (b) We calculate CIDEr for both specialised and generalist models on different datasets.

X-RGen closely emulates the behaviour of human radiologists, which we have distilled into four key phases: 1) initial observation, 2) cross-region analysis, 3) medical interpretation, and 4) report formation. Firstly, akin to a radiologist’s initial assessment of medical images, X-RGen adopts an image encoder to identify crucial features within radiology images in the **initial observation** phase. Then, much like how radiologists deepen their understanding of a patient’s health by drawing upon their extensive experience with past cases, X-RGen aims to enhance the recognition ability of the image encoder by leveraging comparisons across images and reports from various anatomical regions in the **cross-region analysis** phase. After that, mirroring radiologists’ use of their expertise for image interpretation, our model similarly employs pre-defined radiological knowledge to conduct an in-depth clinical analysis of the enhanced features for **medical interpretation**. In this way, the model would focus more on medical-relevant terms, even those infrequent in the training corpus. Finally, in the **report formation** phase, X-RGen compiles these insights into coherent and detailed reports.

We conduct experiments on a merged dataset, covering six different anatomical regions, *i.e.*, chest, abdomen, knee, hip, wrist and shoulder. For a fair comparison, we include chest images from a widely used public dataset – IU-Xray [7]. For the other five regions, we use our private data. To evaluate the performance, we apply the natural language generation metrics (BLEU [31] and CIDEr [34]) and clinical efficacy metrics (recall and F1 score [26]). The results (see Figure 2b) show the superiority of X-RGen compared with both specialised (trained on each single dataset) and generalist models (trained on the merged dataset).

In summary, our contributions include:

- We propose X-RGen, a framework inspired by the behaviour of radiologists for generating reports across various anatomical regions. This framework contains four main phases: initial observation, cross-region analysis, medical interpretation, and report formation.

- We enhance image recognition through cross-region analysis (CA), improving alignment between images and reports across anatomical areas. In medical interpretation (MI), we integrate radiology-specific knowledge, alleviating the ignoring of rare yet crucial terms during report generation.
- We verify the superiority of our X-RGen on seven datasets w.r.t. different anatomical regions. The experimental results on both NLG and clinical efficacy metrics demonstrate the effectiveness of the proposed X-RGen.

## 2 Related Works

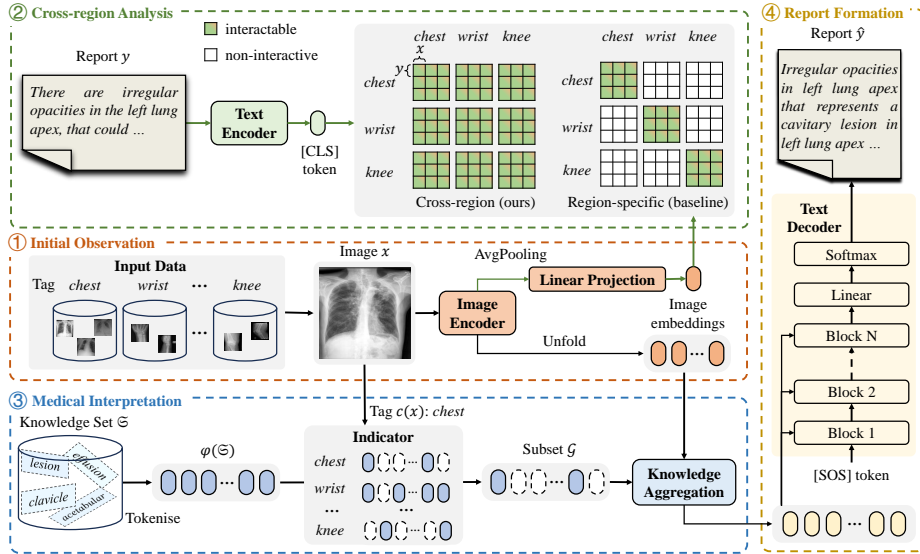
**Image Captioning** Natural image captioning [1, 13, 35, 43] seeks to automatically generate descriptive captions for given images. Many methods [6, 27, 30, 32] have been proposed, leading to significant advancements in the state-of-the-art. The typical image captioning models [18, 35] mainly contain two parts: a CNN-based image encoder and an RNN-based decoder for generating captions. Several studies [13, 50] have incorporated the attention mechanism [33] into the diagram, encouraging the models to pay greater attention to the highlighted regions. However, radiology report generation requires specialised knowledge of medical imaging and terminology, while natural image captioning is more general.

**Radiology Report Generation** Radiology report generation focuses on medical imaging data to produce detailed and accurate reports that encapsulate findings, interpretations, and diagnoses from medical images. Previous works [17, 40, 44, 47] employ a hierarchical LSTM for the long paragraph generation in medical reports. To further enhance performance, several studies [4, 5, 12, 21, 37, 38] adopt a Transformer as the report decoder, leading to notable improvements in results. Moreover, to capture the radiology terminologies and their semantic relationships, recent works [22, 24, 25, 49] take knowledge graphs as inputs or optimisation constraints (*e.g.*, classification labels). However, these models are based on a single-dataset training-testing paradigm, while radiologists often write reports according to radiology images w.r.t. various body regions. When these models are applied directly to another dataset that contains different body regions, they often encounter significant performance degradation issues.

While other knowledge-based models like [24, 46] develop a knowledge graph, the relations (edges) between topics (nodes) cannot be updated during training, which limits its effectiveness for exploiting implicit relationships. Besides, this graph focuses on chest X-rays only, restricting its applicability to other anatomical regions. Thus, we tend to reorganise the topics in our knowledge set such that they cover the medical terminologies relevant to a broad range of body regions without predefined and fixed relations, where the relations are learnable.

## 3 Method

Our X-RGen (see Figure 3) contains four phases: 1) initial observation, 2) cross-region analysis, 3) medical interpretation and 4) report formation. We first use



**Fig. 3:** Overall of X-RGen. We decompose the framework into four phases: 1) initial observation, 2) cross-region analysis, 3) medical interpretation, and 4) report formation. Specifically, starting with an image encoder to extract visual features, the model then enhances recognition by interacting with cross-region data. Next, it applies radiological knowledge for further medical-aware analysis, and finally, generates reports based on the enhanced and medical-aware features. Note that the second phase (*i.e.*, cross-region analysis, **green** arrows) is only for training and will be removed in inference.

an image encoder  $f$  for feature extraction from images  $x$ . Then, we boost the image feature  $\mathcal{O}$  to  $\hat{\mathcal{O}}$  during training by improving recognition ability through cross-region data interaction. Subsequently, we integrate radiological knowledge for a deeper medical interpretation and generate medically enhanced features  $\mathcal{Z}$ . Last, we yield a radiology report  $y$  from  $\mathcal{Z}$ . Notably, the cross-region analysis phase (indicated by green arrows in Figure 3) is excluded in inference.

### 3.1 Initial Observation

We introduce a CNN-based image encoder  $f$  to simulate the initial examination phase of radiologists analysing radiology images. This encoder processes the input image  $x$  using convolutional layers to extract feature maps, which capture crucial diagnostic information. The extracted features are unfolded through a **Unfold** operation, creating a set of feature embeddings. This CNN architecture dynamically focuses on important image details, mirroring a radiologist’s method of identifying key features. Mathematically,

$$\mathcal{O} = \text{Unfold}(f(x)), \quad (1)$$

where  $\mathcal{O}$  is the visual embeddings, which can be defined as  $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$ . By simulating the initial observation phase, the image encoder can effectively

capture and prioritise the most relevant features within the radiology images before proceeding to more detailed analysis and interpretation phases. This approach ensures that the encoder, much like a radiologist, establishes a global understanding of the image, setting a solid foundation for accurate medical interpretation and subsequent report generation.

### 3.2 Cross-region Analysis

With the above observation, radiologists develop a further understanding of the patient’s health status based on their experience, which is learned from reviewing and analysing numerous past cases. Similarly, in our cross-region analysis phase, we seek to improve the recognition ability of the image encoder  $f$  by analysing images and reports from different anatomical regions.

**Unified Representation** We first adopt the image encoder  $f(x)$  to extract visual features from multi-region images, where  $x$  represents images from various anatomical regions. Then, the extracted features are summarised through an average pooling (**AvgPooling**) followed by a linear projection layer<sup>3</sup>, creating a comprehensive image embedding with a specific dimension. For reports, we use a Transformer-based text encoder  $g$  to process the corresponding reports  $y$ , *i.e.*,  $\mathcal{W} = g(y)$ , where  $\mathcal{W} = \{w_1, w_2, \dots, w_m, w_{[\text{CLS}]}\}$  is the set of word tokens.

**Enhancing Recognition with Cross-region Learning** For a more comprehensive understanding of the human body, we seek to enable models to consider how different anatomical regions relate to each other. From Figure 3, previous region-specific works [16, 22] focus only on alignments among images and reports within the same anatomical region. Unlike these, we adapt the learning objective for cross-region analysis, allowing for interactions across different anatomical regions. We encode images and reports with  $\zeta(f(x))$  and  $g(y)$ , respectively, into the shared space. Formally, the cross-region learning objective can be defined as

$$\begin{aligned} \mathcal{L}_{i2r} &= -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{\exp\left(\sigma(\zeta(f(x_i)), g(y_i))\right)}{\sum_{j=1, i \neq j}^{|\mathcal{B}|} \exp\left(\sigma(\zeta(f(x_i)), g(y_j))\right)}, \\ \mathcal{L}_{r2i} &= -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{\exp\left(\sigma(g(y_i), \zeta(f(x_i)))\right)}{\sum_{j=1, i \neq j}^{|\mathcal{B}|} \exp\left(\sigma(g(y_i), \zeta(f(x_j)))\right)}, \\ \mathcal{L}_x &= \frac{1}{2}(\mathcal{L}_{i2r} + \mathcal{L}_{r2i}), \end{aligned} \quad (2)$$

where  $\sigma$  is the similarity function that calculates the cosine similarity between  $\zeta(f(x))$  and  $w_{[\text{CLS}]} \in \mathcal{W} = g(y)$ . The sum in the denominator runs over all

<sup>3</sup> For simplicity, we represent the whole process as  $\zeta$ , including both **AvgPooling** and linear projection.

image-report pairs  $(x, y)$  within the mini-batch  $\mathcal{B}$ . Notably, the pairs selected within each mini-batch span multiple anatomical regions, which ensures diversity and holistic learning across different body regions. In this way, the model gains a deeper, more generalised insight into the semantics across varied regions, improving its recognition capacity. In general, the image embeddings before and after enhancement can be represented as

$$\begin{aligned} \mathcal{T} &:= \mathcal{O} \rightarrow \tilde{\mathcal{O}} \\ &:= \{o_1, o_2, \dots, o_n\} \rightarrow \{\tilde{o}_1, \tilde{o}_2, \dots, \tilde{o}_n\}, \end{aligned} \quad (3)$$

where  $\mathcal{T}$  refers to the process of our cross-region analysis during training.  $\tilde{\mathcal{O}}$  is the enhanced image embeddings from the input image  $x$ .

### 3.3 Medical Interpretation

**Building General Radiological Knowledge Set** We seek to construct a general knowledge set  $\mathfrak{S}$  that covers the most common abnormalities or findings in the radiology reports. For convenience, we call each item in this knowledge set a topic like [24, 49]. While they develop a knowledge graph, the relations (edges) between topics (nodes) cannot be updated during training, which limits its effectiveness for exploiting implicit relationships. Furthermore, this knowledge graph focuses on chest X-rays only, which restricts its applicability to other body parts and broader use cases. Thus, we have reorganised the topics in our knowledge set such that they cover the medical terminologies relevant to a broad range of body parts without pre-defined and fixed relations.

To better build a generic knowledge base, following [42], we adopt one of the most useful natural language processing methods, called topic modelling, on the database, that seeks to characterise the knowledge for each body part with a series of topics namely  $\mathcal{G}$ . Specifically, we first use spacy [29], currently the most popular entity detection tool, to extract the medical entities and obtain the 50 most frequent words for each body part. Then, 20  $\sim$  30 most critical words are further filtered by radiologists to create the existing knowledge base. Mathematically, the general set  $\mathfrak{S}$  can be defined as  $\mathfrak{S} = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \dots$ .

**Region-aware Knowledge Selection** Understanding the context and clinical nuances in radiology reports requires deep medical knowledge and expertise. Models may lack the comprehensive understanding needed to generate reports that incorporate relevant clinical information, leading to inaccuracies or missing crucial details. To address this, we introduce a condition signal  $c$  to selectively activate a set of topics based on the anatomical regions associated with the input image. By doing so, our model is able to filter out topics that are irrelevant to the specific region before conducting reasoning between the given image  $x$  and the general knowledge set  $\mathfrak{S}$ . Specifically, we devise an indication function  $\mathbb{1}(\cdot)$ , which enables the selection of the topics in  $\mathfrak{S}$ . Formally,

$$\mathcal{G} = \mathbb{1}(\varphi(\mathfrak{S})|c(x)), \quad (4)$$

where  $\varphi$  is the pre-trained tokeniser for word embeddings while  $\mathcal{G}$  is the subset of the general knowledge  $\mathfrak{S}$  (*i.e.*,  $\mathcal{G} \subseteq \mathfrak{S}$ ), denoting the selected topics. Here,  $c(x)$  is a tag of the body region the given image  $x$  belongs to (*e.g.*, “*chest*”), which is manually predefined in advance<sup>4</sup>. For instance, if the tag of an input image is predefined as “*chest*”, we choose a specific set of topics – namely, “*airspace disease, atelectasis, calcinosis, cardiomegaly, cicatrix, edema, effusion, emphysema, fractures, hernia, hypoinflation, lesion, medical device, normal, opacity, other, pneumonia, pneumothorax, scoliosis, thickening*” – to represent the knowledge base for this image. Notably, the topics selected for different samples belonging to the same body region are identical. In our work, we define six different tags, including *chest, abdomen, knee, hip, wrist, and shoulder*.

**Intra-region Knowledge Aggregation** For knowledge aggregation, our idea is to design a learnable aggregation model  $\pi$  with a capacity for cross-modal reasoning, allowing it to determine the most relevant topics between the knowledge set  $\mathcal{G}$  and the enhanced embeddings of image  $\tilde{\mathcal{O}}$ . A straightforward way is adopting scaled dot product attention, which enables the topics and images to interact with one another. Concretely, we design a knowledge-image co-attention module using a  $l$ -layer Transformer. Formally,

$$\mathcal{Z} = \pi(\tilde{\mathcal{O}}, \mathcal{G}) = \text{Transformer}([\tilde{\mathcal{O}}; \mathcal{G}]), \quad (5)$$

Here,  $[\cdot; \cdot]$  is the concatenation operation.  $\mathcal{Z}$  is the set of aggregated embeddings, *i.e.*,  $\mathcal{Z} = \{z_1, z_2, \dots, z_n, \dots, z_{n+k}\}$ , where  $k$  is the number of topics in  $\mathcal{G}$ .

### 3.4 Report Formation

**Text Decoder** Based on the aggregated embeddings  $\mathcal{Z}$ , we adopt a Transformer-based text decoder, containing  $N$  Transformer blocks, for generating the final report (see Figure 3). Concretely, the decoding process starts by feeding a special start token [SOS] to our text decoder, along with positional embeddings. The decoder uses a self-attention mechanism to process the start token, positional embeddings, and aggregated embeddings. The whole process can be defined as

$$\hat{y} = h(\mathcal{Z}) = \arg \max \prod_{t=1}^T p(\hat{w}_t | \hat{w}_{i < t}, \mathcal{Z}), \quad (6)$$

where  $h$  refers to the text decoder.  $\hat{y}$  is the generated report and  $\hat{w}_t$  is the  $t$ -th predicted word, *i.e.*,  $\hat{w}_t \in \hat{y}$ . The decoder generates the report auto-regressively, attending to the aggregated tokens and previously generated words at each step. In each step, it applies a softmax function to predict the next word’s probability distribution over the entire vocabulary. The process is repeated until an end token is generated or a predefined maximum sequence length  $T$  is reached.

<sup>4</sup> In clinical practice, since doctors specify which specific body part to image before taking medical images, the corresponding part naturally has a tag.



---

**Algorithm 1** Overall Algorithm for X-RGen.

---

**Require:** Training triplets  $\{x, c(x), y\}$  w.r.t. image, tag, and report; X-RGen with modules: image encoder  $f$ , project layer  $\zeta$ , text encoder  $g$ , aggregation model  $\pi$ , text decoder  $h$ .

- 1: Construct general knowledge set  $\mathfrak{S}$  across multiple anatomical regions.
- 2: // *Training*
- 3: **while** *not convergent* **do**
- 4:   Extract image embeddings  $\mathcal{O}$  from each input image  $x$  with Eq. (1).
- 5:   Boost  $\mathcal{O}$  to  $\tilde{\mathcal{O}}$  by updating  $f$ ,  $\zeta$  and  $g$  using the objective in Eq. (2).
- 6:   Select region-aware knowledge  $\mathcal{G}$  from  $\mathfrak{S}$  according to tag  $c(x)$  in Eq. (4).
- 7:   Obtain medical-aware image tokens  $\mathcal{Z}$  by aggregating  $\mathcal{G}$  and  $\tilde{\mathcal{O}}$  with Eq. (5).
- 8:   Generate report  $\hat{y}$  from  $\mathcal{Z}$  with Eq. (6).
- 9:   Update  $f$ ,  $\zeta$ ,  $g$ ,  $\pi$ , and  $h$  by minimising the objective in Eq. (7).
- 10: **end while**
- 11: // *Inference*
- 12: Extract embeddings  $\tilde{\mathcal{O}}$  from image  $x$  using  $f$  with **Unfold** operation in Eq. (1).
- 13: Select knowledge  $\mathcal{G}$  by Eq. (4) and then aggregate it with  $\tilde{\mathcal{O}}$  by Eq. (5) to get  $\mathcal{Z}$ .
- 14: Generate report  $\hat{y}$  from  $\mathcal{Z}$  with Eq. (6).

---

### 3.5 Training and Inference

**Overall Training Objective** As shown in Algorithm 1, our overall training objective contains a captioning loss  $\mathcal{L}_{cap}$  and the cross-region loss  $\mathcal{L}_x$ <sup>5</sup>, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{cap} + \lambda \mathcal{L}_x, \quad (7)$$

where  $\lambda$  is a hyper-parameter to balance these two terms. Typically, sequence generation models are trained using the autoregressive Teacher Forcing scheme, to maximise the probability of the ground-truth token  $w_t$  given all previous ground-truth tokens  $w_{i < t}$ . The captioning loss function can be formulated as

$$\mathcal{L}_{cap}(x, y) = -\log p(y|x) = \sum_{t=1}^T -\log p(w_t | w_{i < t}, x), \quad (8)$$

where  $w_t$  is the  $t$ -th token in report  $y$ , and  $T$  is the total number of words in  $y$ .

**Inference** As shown in Algorithm 1, given a radiology image  $x$ , we use the image encoder  $f$  to extract image features and use **Unfold** operation to obtain a set of image embeddings  $\tilde{\mathcal{O}}$ . After that, we select region-relevant knowledge  $\mathcal{G}$  from the general knowledge set  $\mathfrak{S}$  based on tag  $c(x)$  and then aggregate  $\mathcal{G}$  with image embeddings  $\tilde{\mathcal{O}}$  to obtain the medical-aware features  $\mathcal{Z}$ . Last, we generate the report  $\hat{y}$  from  $\mathcal{Z}$ .

Note that in inference, instead of relying on previous ground-truth word tokens, we predict the next word token based on the tokens that have been previously predicted in an auto-regressive manner. Besides, the cross-region analysis

<sup>5</sup>  $\mathcal{L}_x$  is the same as Eq. (2) in Section 3.2.

is dropped in inference since its primary role is to boost the image encoder’s recognition capabilities during training through the cross-region alignment loss.

## 4 Experiments and Results

### 4.1 Datasets

In experiments, we construct a merged dataset that contains paired data w.r.t. six anatomical regions, including chest, abdomen, knee, hip, wrist and shoulder. Due to the lack of existing datasets, we collect private image-report pairs on all six anatomical regions. Anonymous Human Research Ethics Committee provides ethics approval for private data used in this study. For each region, we have 3,000 patients and the ratio of train/val/test is 70%/15%/15%. Notably, for a fair comparison with previous works, we use chest pairs on IU-Xray [7], a publicly recognised dataset, rather than our private ones. It consists of 3,955 fully de-identified radiology reports, each paired with frontal and/or lateral chest X-ray images. Following [5, 22], we remove cases that contain only a single image and then divide the dataset into train, validation, and test sets with 2069/296/590 pairs, respectively.

### 4.2 Evaluation Metrics and Implementation Details

To assess the quality of generated reports, we adopt widely used natural language generation (NLG) metrics, *i.e.*, BLEU (B1~B4) [31], ROUGE [23], METEOR [2] and CIDEr [34]. We assess the clinical efficacy of generated reports using recall and F1 score [26] along with a CLIP-based metric, called CLIPScore [11]. The CLIPScore<sup>6</sup> can assess whether the generated reports are semantically aligned with given images, even when they are different from the reference reports.

To demonstrate the enhanced capacity for semantic understanding offered by the image encoder, we undertake the linear classification probing evaluation using the CheXpert [14] dataset, which contains five individual binary labels: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. For this process, we fix the image encoder, which has been trained on our X-RGen, and exclusively train a randomly initialised linear classification head.

We use ResNet101 [10], pre-trained on ImageNet [8], as image encoder. We use the tokeniser and text encoder from MedClip [41] to convert words to embeddings. The knowledge aggregation module consists of a three-layer Transformer [9]. We resize input images to  $224 \times 224$ , and limit the maximum epochs to 100 and use Adam [19] with a weight decay of  $1e-4$ . We set the  $\lambda$  to 1.0.

### 4.3 Comparison with State-of-the-arts

**Specialised Baselines** We compare X-RGen with the existing report generation methods, including R2Gen [5], R2GenCMN [4], MSAT [39], DCL [22] and

<sup>6</sup> We use MedClip [41] instead of the original CLIP trained on the natural domain.

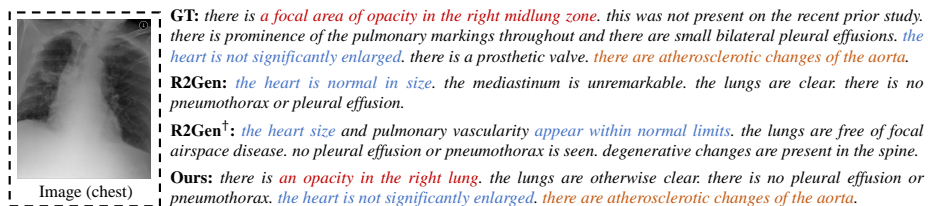
	Chest		Abdomen		Knee		Hip		Wrist		Shoulder		Ave	
	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr	B4	CIDEr
specialized models														
Transformer [33]	0.162	0.511	<u>0.108</u>	0.261	0.079	0.151	0.077	0.137	0.086	0.129	0.088	0.192	0.100	0.230
R2Gen [5]	0.165	0.430	0.105	0.248	0.077	<u>0.193</u>	0.082	0.210	<u>0.093</u>	<u>0.288</u>	0.082	0.174	0.101	0.257
R2GenCMN [4]	0.170	<u>0.641</u>	0.102	0.161	0.083	0.164	0.083	0.220	0.087	0.212	0.082	0.134	0.101	0.255
MSAT [39]	0.171	0.394	0.105	<u>0.275</u>	0.082	0.135	0.081	<u>0.235</u>	0.081	0.180	0.080	0.173	0.100	0.232
DCL [22]	0.163	0.586	-	-	-	-	-	-	-	-	-	-	-	-
METransformer [38]	<u>0.172</u>	0.435	-	-	-	-	-	-	-	-	-	-	-	-
X-RGen (ours)	0.163	0.609	0.106	0.196	<u>0.087</u>	0.175	<u>0.086</u>	0.192	0.089	0.243	<u>0.088</u>	<u>0.197</u>	<u>0.103</u>	<u>0.269</u>
generalist models														
R2Gen <sup>†</sup> (bs=16)	0.084	0.289	0.104	0.280	0.064	0.154	0.074	0.203	0.085	0.217	0.082	0.186	0.082	0.222
R2Gen <sup>†</sup> (bs=96)	0.147	0.470	0.097	0.271	0.075	0.181	0.080	0.226	0.084	0.258	0.095	0.274	0.096	0.280
R2Gen <sup>†</sup> (bs=192)	0.114	0.359	0.100	0.271	0.089	0.204	<b>0.086</b>	0.238	<b>0.102</b>	0.296	0.096	0.277	0.098	0.274
X-RGen (ours, bs=16)	0.152	0.509	0.108	0.276	0.071	0.166	0.073	0.184	0.079	0.229	0.084	0.220	0.095	0.264
X-RGen (ours, bs=96)	0.161	<b>0.700</b>	0.110	0.292	0.077	0.188	0.084	<b>0.257</b>	0.090	0.255	<b>0.099</b>	0.272	0.104	0.327
X-RGen (ours, bs=192)	<b>0.177</b>	0.602	<b>0.118</b>	<b>0.327</b>	<b>0.093</b>	<b>0.242</b>	0.076	0.215	0.097	<b>0.305</b>	0.096	<b>0.287</b>	<b>0.110</b>	<b>0.330</b>
	Chest		Abdomen		Knee		Hip		Wrist		Shoulder		Ave	
	F	R	F	R	F	R	F	R	F	R	F	R	F	R
specialized models														
Transformer [33]	0.584	0.624	0.559	0.546	0.486	<u>0.464</u>	0.525	0.481	0.506	0.453	0.463	0.420	0.521	0.498
R2Gen [5]	0.583	<b>0.655</b>	0.558	0.554	0.462	0.389	0.496	0.427	0.514	0.479	<u>0.520</u>	0.468	0.522	0.495
R2GenCMN [4]	0.592	0.645	0.540	0.505	0.491	0.437	<u>0.528</u>	0.501	0.500	0.427	0.462	0.387	0.484	0.519
X-RGen (ours)	<u>0.593</u>	0.642	<u>0.565</u>	<u>0.559</u>	<u>0.497</u>	0.460	0.522	<b>0.502</b>	<u>0.533</u>	<u>0.506</u>	0.508	<u>0.474</u>	<u>0.536</u>	<u>0.524</u>
generalist models														
R2Gen <sup>†</sup> (bs=192)	0.589	0.578	0.561	0.549	0.495	0.443	0.512	0.496	0.531	0.496	0.505	0.479	0.532	0.507
X-RGen (ours, bs=192)	<b>0.594</b>	0.647	<b>0.580</b>	<b>0.565</b>	<b>0.501</b>	<b>0.467</b>	<b>0.529</b>	0.499	<b>0.543</b>	<b>0.512</b>	<b>0.514</b>	<b>0.482</b>	<b>0.544</b>	<b>0.529</b>

**Table 1:** Comparison of NLG metrics (upper: B4 and CIDEr) and clinical efficacy metrics (lower: F  $\rightarrow$  F1 ; R  $\rightarrow$  recall) with the recent specialised models on six datasets. <sup>†</sup> means we optimise the model on our merged training dataset while the “bs” is the training batch size. All evaluations are conducted on the test set. A higher value means better performance. We highlight the best results on specialised models with underline while the best results on all models (both specialised and generalist) with **bold**.

METransformer [38]. Besides, we consider a widely used natural image captioning method (*i.e.*, Transformer [33]) as another baseline. First, we individually optimise our model and each baseline in a specialised training setting. For a fair comparison, we adopt the batch size (bs) of 16, which is a commonly used setting in the report generation task<sup>7</sup>. In Table 1, compared with specialised baselines, our X-RGen achieves superior results in both NLG (average B4 and CIDEr) and clinical efficacy metrics (average F1 and recall scores). This indicates that the radiologist-minded framework benefits even in the specialised setting.

**Generalist Baselines** To further analyse the performance of X-RGen, we adapt specialised models into the joint training setting due to the lack of existing generalist baselines. Here, we use all the training data on different subsets for optimisation. To mitigate the impact of different architectures, we select R2Gen [5] as the baseline. The main difference between R2Gen and our base model lies

<sup>7</sup> We also experiment with increasing the batch size of the baselines to improve their performance, but it only results in performance comparable to bs = 16.



**Fig. 4:** Reports generated by X-RGen (ours) and two baselines – R2Gen and R2Gen<sup>†</sup>. R2Gen is trained on IU-Xray only while R2Gen<sup>†</sup> optimised on our merged training set.

in the text decoder, where R2Gen has an additional Relational Memory (RM) module while our model does not include it. For a fair comparison, we adjust the batch size (bs) to match our setting. Specifically, we increase it from 16 to 96 and 192, which aligns with our own configuration, thereby mitigating the potential performance improvement attributed solely to the larger batch size.

Table 1 shows that regardless of  $bs = 96$  or  $192$ , our X-RGen consistently outperforms R2Gen in terms of both average B4 and CIDEr scores, which demonstrates its effectiveness in generating accurate and high-quality radiology reports. Moreover, R2Gen (generalist) has an  $\sim 9\%$  improvement in CIDEr (0.257 to 0.280) while achieving a comparable result in B4 (0.101 and 0.098) compared with R2Gen (specialised). This indicates the positive impact of using diverse and increased training data. For our X-RGen, the generalist version achieves larger improvements in both CIDEr ( $\sim 22\%$ : 0.269 to 0.330) and B4 ( $\sim 7\%$ : 0.103 to 0.110) compared with the specialised counterpart. A similar phenomenon also occurs in clinical efficacy metrics (*i.e.*, average F1 and recall scores) in Table 1. These results demonstrate that the gains in performance are not solely attributed to the dataset, but also due to the benefits provided by the proposed radiologist-minded framework.

#### 4.4 Qualitative Evaluation

In this part, we further assess the quality of reports generated by different methods, including our method and two baselines, *i.e.*, R2Gen and R2Gen<sup>†</sup>, trained on IU-Xray (chest) only and our merged dataset, respectively. In Figure 4, we highlight the descriptions in different colours (red, blue and orange), which are semantically aligned with those in the ground-truth (GT) reports. When considering the prominent area (*e.g.*, the heart), all three models can provide (almost) accurate descriptions. However, R2Gen still tends to generate the “average” descriptions like “*the heart is normal in size*” while R2Gen<sup>†</sup> shows improvement due to optimisation with a more diverse dataset. Moreover, our X-RGen shows a more powerful capacity to generate descriptions that align semantically with the ground truth. For instance, while the baselines fail to capture the details, our model accurately describes “*an opacity in the right lung*”, matching the GT description: “*a focal area of opacity in the right midlung zone*”.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Base	0.412	0.255	0.175	0.129	0.176	0.340	0.426
+MI*	0.414	0.263	0.184	0.137	0.180	0.342	0.474
+MI	<b>0.457</b>	0.284	0.204	0.156	0.182	0.349	0.537
+CA	0.454	<b>0.290</b>	<b>0.210</b>	<b>0.161</b>	<b>0.187</b>	<b>0.361</b>	<b>0.700</b>

**Table 2:** Performance analysis on Chest (IU-Xray). “Base” contains only the initial observation and report formation phases. Both the medical interpretation (MI) phase and MI\* are applied on top of the Base, where the MI\* means MI without indicator  $\mathbb{1}(\cdot)$  in Eq. (4). The cross-region analysis (CA) is applied only on top of the Base+MI.

#### 4.5 Ablation Study

In this part, we evaluate the performance of our base model with and without the medical interpretation (MI) and cross-region analysis (CA) phases on Chest (IU-Xray). In Table 2, the results show that our base model with MI alone achieves better performance compared with the counterpart without it (*i.e.*, B4: 0.129  $\rightarrow$  0.156 while CIDEr: 0.426  $\rightarrow$  0.537), which verifies the significance of the radiology-relevant knowledge in report generation task. While MI\*, without the indicator  $\mathbb{1}(\cdot)$ , can also achieve improved results compared to the base model (*e.g.*, B4: 0.129  $\rightarrow$  0.137), it is surpassed by MI (B4: 0.156). This highlights the importance of region-specific guidance and demonstrates the necessity of incorporating such guidance for better performance. Finally, incorporating the CA phase further enhances the performance, resulting in the best scores for both B4 (0.161) and CIDEr (0.700). This demonstrates the contribution of the CA in improving the model performance by leveraging guidance from different modalities, including both images and reports.

#### 4.6 Discussions

In this part, we explore how well our X-RGen aligns semantically between images and reports. We also assess the impact of our cross-region analysis (CA) and medical interpretation (MI) on this semantic alignment. Besides, we evaluate the recognition capacity of our image encoder by linear probing on CheXpert.

**Semantic Alignment between Image and Report** Besides the reference-based metrics like BLEU4, which may be influenced by semantically irrelevant factors (*e.g.*, writing style [3]), we seek to directly assess the semantic alignment between the input images and the generated reports. Thus, we calculate a reference-free score, namely CLIPScore [11], for R2Gen [5] (trained on IU-Xray only), R2Gen<sup>†</sup> (trained on the merged dataset) and our X-RGen. Notably, as we use MedClip [41] in CLIPScore, which is pre-trained on chest X-ray datasets, we only evaluate this score on the IU-Xray (chest) dataset because it is open-sourced. Besides, for a fair comparison, we set the batch size to 96 for both R2Gen<sup>†</sup> (achieves the best results) and our model. For the specialised R2Gen,

CLIPScore		CLIPScore		AUC score	
R2Gen [5]	77.670	Base	75.687	R2Gen [5]	77.435
R2Gen <sup>†</sup> [5]	75.402	+ MI	76.865	R2Gen <sup>†</sup> [5]	79.213
X-RGen (ours)	78.052	+ MI + CA	78.052	X-RGen w/o CA	80.405
<b>(a) Ours vs. R2Gen</b>		<b>(b) Impact of MI and CA</b>		<b>(c) Linear probing</b>	

**Table 3:** We assess (a) semantic alignment between images and reports on IU-Xray (chest), and (b) the effect of medical interpretation (MI) and cross-region analysis (CA) phases for alignment. (c) Linear probing on CheXpert to evaluate the recognition ability of the image encoder. <sup>†</sup> means we optimise the model on our merged dataset.

we keep the settings of the official code unchanged. In Table 3a, our X-RGen outperforms R2Gen with a CLIPScore of 78.052, regardless of whether it is trained on IU-Xray only (77.670) or on our merged dataset (75.402).

Moreover, similarly to Table 3a, we use the CLIPScore to assess whether the model can generate more semantically aligned reports aided by two main phases (*i.e.*, MI and CA). Table 3b reveals that the model with MI produces improved CLIPScore compared to the base counterpart (from 75.687 to 76.865), and incorporating the CA further enhances the performance, resulting in the best CLIPScore. It demonstrates the capability of our MI and CA phases in recognition enhancement, therefore generating more accurate reports.

**Recognition Capacity of Image Encoder** To further investigate the effect of our CA phase in recognition enhancement, we seek to directly test the recognition ability of our image encoder. To this end, we simply add a classification head on top of our image encoder (*i.e.*, linear probing) and then evaluate the performance on a multi-label classification dataset – CheXpert [14]. In Table 3c, our X-RGen obtains an 81.252 AUC score that outperforms both R2Gen (77.435) and R2Gen<sup>†</sup> (79.213), which indicates the ability of our model to correctly recognise and classify different medical diseases within the input radiology images. Moreover, we evaluate the performance of the X-RGen without incorporating CA during training. In this case, the AUC score decreases to 80.405, further demonstrating the effectiveness of CA in enhancing the recognition ability of our model.

## 5 Conclusion

In this paper, we propose X-RGen, a framework designed for automatic radiology report generation across multiple anatomical regions. Unlike previous works, our X-RGen follows the behaviour of human radiologists with four key phases: initial observation, cross-region analysis, medical interpretation, and report formation. The experiments across six X-ray datasets demonstrate the superiority of our X-RGen. Through this work, we hope to mark a step towards narrowing the gap between medical artificial intelligence and human radiologists, starting with a more radiologist-like diagnostic process for the report generation task.

## References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6077–6086 (2018) [4](#)
2. 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) [10](#)
3. Chen, Q., Deng, C., Wu, Q.: Learning distinct and representative modes for image captioning. Adv. Neural Inform. Process. Syst. pp. 9472–9485 (2022) [2](#), [13](#)
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. ACL-IJCNLP pp. 5904–5914 (2022) [4](#), [10](#), [11](#)
5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. EMNLP pp. 1439–1449 (2020) [2](#), [4](#), [10](#), [11](#), [13](#), [14](#)
6. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10578–10587 (2020) [4](#)
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association pp. 304–310 (2016) [3](#), [10](#)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255 (2009) [10](#)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Int. Conf. Learn. Represent. (2021) [10](#)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016) [10](#)
11. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. EMNLP pp. 7514–7528 (2021) [10](#), [13](#)
12. Hou, W., Cheng, Y., Xu, K., Li, W., Liu, J.: Recap: Towards precise radiology report generation via dynamic disease progression reasoning. ACL (2023) [4](#)
13. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Int. Conf. Comput. Vis. pp. 4634–4643 (2019) [4](#)
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. pp. 590–597 (2019) [10](#), [14](#)
15. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods pp. 203–211 (2021) [2](#)
16. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. AAAI (2024) [6](#)
17. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. ACL pp. 2577–2586 (2018) [4](#)

18. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3128–3137 (2015) [4](#)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [10](#)
20. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: Int. Conf. Learn. Represent. (2023) [2](#)
21. Li, M., Cai, W., Verspoor, K., Pan, S., Liang, X., Chang, X.: Cross-modal clinical graph transformer for ophthalmic report generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20656–20665 (2022) [4](#)
22. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3334–3343 (2023) [2](#), [4](#), [6](#), [10](#), [11](#)
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) [10](#)
24. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13753–13762 (2021) [2](#), [4](#), [7](#)
25. Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al.: Auto-encoding knowledge graph for unsupervised medical report generation. Adv. Neural Inform. Process. Syst. pp. 16266–16279 (2021) [2](#), [4](#)
26. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019) [3](#), [10](#)
27. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 375–383 (2017) [4](#)
28. 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. ACL-IJCNLP pp. 269–280 (2021) [2](#)
29. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 319–327. Association for Computational Linguistics, Florence, Italy (Aug 2019). <https://doi.org/10.18653/v1/W19-5034>, <https://www.aclweb.org/anthology/W19-5034> [7](#)
30. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10971–10980 (2020) [4](#)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002) [3](#), [10](#)
32. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7008–7024 (2017) [4](#)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017) [4](#), [11](#)
34. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4566–4575 (2015) [3](#), [10](#)
35. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3156–3164 (2015) [4](#)



36. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Adv. Neural Inform. Process. Syst.* pp. 33536–33549 (2022) [2](#)
37. Wang, Z., Han, H., Wang, L., Li, X., Zhou, L.: Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging* pp. 2803–2813 (2022) [4](#)
38. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 11558–11567 (2023) [2](#), [4](#), [11](#)
39. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: *MICCAI*. pp. 655–664 (2022) [10](#), [11](#)
40. Wang, Z., Zhou, L., Wang, L., Li, X.: A self-boosting framework for automated radiographic report generation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2433–2442 (2021) [4](#)
41. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163* (2022) [10](#), [13](#)
42. Willeminck, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. *Radiology* pp. 4–15 (2020) [7](#)
43. 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: *ICML*. pp. 2048–2057 (2015) [4](#)
44. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: *MICCAI*. pp. 457–466 (2018) [4](#)
45. Yan, S., Cheung, W.K., Chiu, K., Tong, T.M., Cheung, K.C., See, S.: Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging* (2023) [2](#)
46. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis* p. 102510 (2022) [4](#)
47. Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: *MICCAI*. pp. 721–729 (2019) [4](#)
48. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 1195–1204 (2021) [2](#)
49. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: *AAAI*. pp. 12910–12917 (2020) [2](#), [4](#), [7](#)
50. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: *AAAI*. pp. 13041–13049 (2020) [4](#)