Complex Organ Mask Guided Radiology Report Generation

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, Weidong Cai
University of Sydney, Sydney, Australia
{*tigu*8498, *zhli*0736}@uni.sydney.edu.au, {*dongnan.liu*, *tom.cai*}@sydney.edu.au

Abstract

The goal of automatic report generation is to generate a clinically accurate and coherent phrase from a single given X-ray image, which could alleviate the workload of traditional radiology reporting. However, in a real-world scenario, radiologists frequently face the challenge of producing extensive reports derived from numerous medical images, thereby medical report generation from multi-image perspective is needed. In this paper, we propose the ***C****omplex* ***O****rgan* ***M****ask* ***G****uided (termed as COMG) report generation model, which incorporates masks from multiple organs (e.g., bones, lungs, heart, and mediastinum), to provide more detailed information and guide the model's attention to these crucial body regions. Specifically, we leverage prior knowledge of the disease corresponding to each organ in the fusion process to enhance the disease identification phase during the report generation process. Additionally, cosine similarity loss is introduced as target function to ensure the convergence of cross-modal consistency and facilitate model optimization. Experimental results on two public datasets show that COMG achieves a* 11.4% *and* 9.7% *improvement in terms of BLEU@4 scores over the SOTA model KiUT on IU-Xray and MIMIC, respectively. The code is publicly available at* https://github.com/GaryGuTC/COMG_model.

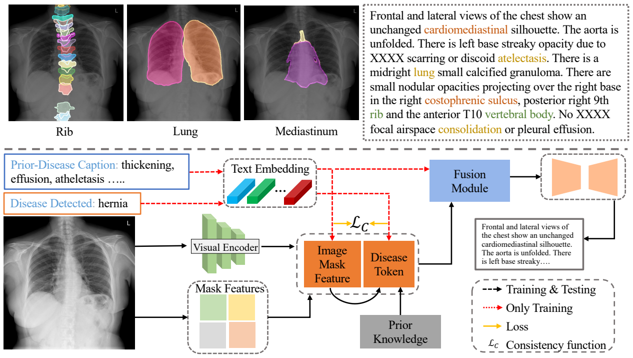

Rib
Lung
Mediastinum
Frontal and lateral views of the chest show an unchanged cardiomediastinal silhouette. The aorta is unfolded. There is left base streaky opacity due to XXXX scarring or discoid atelectasis. There is a midright lung small calcified granuloma. There are small nodular opacities projecting over the right base in the right costophrenic sulcus, posterior right 9th rib and the anterior T10 vertebral body. No XXXX focal airspace consolidation or pleural effusion.
Prior-Disease Caption: thickening, effusion, atheletasis .....
Disease Detected: hernia
Text Embedding
...
Visual Encoder
Mask Features
Image Mask Feature
Disease Token
$\mathcal{L}_C$
Prior Knowledge
Fusion Module
Frontal and lateral views of the chest show an unchanged cardiomediastinal silhouette. The aorta is unfolded. There is left base streaky....
Training & Testing
Only Training
Loss
$\mathcal{L}_C$ Consistency function

Figure 1. Top: the above section display the relationship between the mask images and the captions, that includes findings related to bone (green), lung (yellow), and mediastinum (red). Bottom: the below section illustrates the basic structure of the COMG model. Best viewed by zooming in.

1. Introduction

Radiology image analysis plays a pivotal role in disease detection [40]. In clinical practice, it is time-consuming, costly, and error-prone for radiologists to review numerous radiology images and generate the corresponding reports for further analysis. To mitigate this challenge, there has been a growing interest in exploring automated radiology report generation (RRG) techniques [26, 31, 38]. Recently, deep learning methods have been widely explored to generate a textual analysis for given images (e.g., image captioning) [1, 7, 44]. However, RRG differs from the standard image caption task and is more complex and challenging [39]. Radiology images focus on some specific regions related to the disease, which account for only a small fraction of the entire image [21]. The main difference between radiographs is the specific area associated with the disease. Such challenges also remain in the text descriptions for RRG, i.e., the differences are mainly from the analysis of the diseases, while the descriptions for the normal tissues are similar.

Several RRG techniques have been proposed to solve the challenges from various perspectives and achieve appealing performance. Specifically, [5, 6, 30, 39] have been developed by improving the encoder-decoder structure and enhancing the feature fusion across different modalities. By incorporating prior knowledge from auxiliary resources, such as the region detection prediction, and retrieval of textual information, [21, 35, 36] are proposed to further enhance the report generation via the comprehensive knowledge. Despite their outstanding performance, none of these methods consider the pixel-level information for the specific tissues. As indicated in [32], the pixel-level segmentation masks for the tissues are also critical for further analysis in radiology images, which reflects the semantic correlations of the tissue, as indicated in Fig. 1. For example, the masks can indicate the size and shape of the tissue, as well as their spatial distributions, which are related to the recognition and understanding of diseases [10]. Since such critical information has not been considered for RRG, existing methods are suboptimal by missing such pixel-level semantic tissue correlations for radiology images.

In this work, we propose a **C**omplex **O**rgan **M**ask **G**uided (COMG) framework for radiology report generation which considers the dense pixel-level information of X-ray images. Specifically, pre-trained segmentation models (e.g., CXAS [32]) are employed to extract segmentation masks for key tissues correlated to the disease diagnosis in the generated reports (e.g., heart, bones, etc.), as indicated in Fig. 1. The mask information provides the model with local-level semantic information on the organs/tissues during report generation, including the morphological structures of specific tissues, and the spatial relationships among different tissues. Then, the feature prototypes are extracted according to the masks for different organs. In addition, we further incorporate the disease keywords associated with each specific tissue as text-level guidance, whose feature embeddings are fused with the feature prototypes to further enhance the understanding of the correlations between each specific tissue and the corresponding disease during the report generation learning. Finally, cross-modal consistency mechanisms are developed to facilitate feature extraction at the vision and language levels by inducing their similarities.

Our contributions can be summarized as follows: 1) We propose to improve the report generation by enhancing the understanding of the semantic correlations of the tissues in the radiology images via mask information; 2) We introduce the disease keywords associated with each tissue as the text-level prior knowledge to further enhance the tissue feature learning; 3) To facilitate the feature extractions for the image and text, two cross-modal consistency mechanisms are developed; 4) Our proposed COMG method is indicated to be effective by achieving outstanding performance on two public medical report generation benchmarks.

2. Related Works

2.1. Image Caption

Image caption tasks aim to generate text descriptions based on the input images. Early works [3, 19, 26, 28] are developed based on the Long Short-Term Memory (LSTM) model [13] and the Convolutional Neural Network (CNN) model [27]. Recently, the transformer models based on the attention module [37] have been widely employed due to their outstanding ability to process the vision and language features [7, 11, 28, 41]. Despite the appealing performance of these methods in general image caption benchmarks, their applicability is limited for radiology report generation, which is challenging due to the data bias between the normal and disease tissues at the image and text levels [41].

2.2. Radiology Report Generation

In recent years, several deep learning-based methods have been proposed for radiology report generation, which can be mainly categorized into two types: 1) improving the model structure [2,5,6,30,43], and 2) incorporating external modality information [21, 35, 36].

CDGPT2 [2] proposes to fuse visual and semantic features by concatenating them into a multi-modal decoder. AlignTransformer [45] proposed to use the multi-grained transformer (MGT) to improve multi-modal features fusion. In addition, R2Gen [6] and R2GenCMN [5] models are developed to enhance the model structure for filtering and fusing image and caption information separately using the LSTM [13] and CMN modules. R2GenCMM-RL [30] further improves the R2GenCMN via reinforcement learning. XproNet [39] model proposed to initialize the models via prototype matrix initialization, and a multi-label contrastive loss function to guide the optimization process.

Additionally, some works have started to explore leveraging external information to facilitate the report generation process. The RepsNet model [36] utilizes external information from a pre-trained VQA model on the VQA-radiology dataset [18]. It proposes to integrate the features for the answer information from the VQA model with image features for accurate report generation. The RGRG model [35] is developed to first utilize the bounding boxes to detect abnormal regions in the image. Then, some suitable detected areas are chosen for report generation, instead of using the entire image. The DCL model [21] integrates information from a pre-constructed knowledge graph that contains correlations between caption words. Such information is further processed by a dynamic graph encoder and then combined with the image features using a blip-like structure [20] to generate accurate reports via the comprehensive understanding of the disease words. In this work, we propose to incorporate a new type of external knowledge, which is the segmentation masks for the key issues related to the disease mentioned in the text analysis. To the best knowledge, we are making an early attempt to improve the RRG tasks via the pixel-level semantic knowledge for the tissues in the radiology images.

3. COMG

The overview of our proposed **C**omplex **O**rgan **M**ask **G**uided (COMG) method is shown in Fig. 2. Our method is established on the R2GenCMN model [5], which is constructed by an encoder-decoder structure using the transformer model [37] and multi-modal feature fusion mechanism. Our main contributions include the Mask-guided Organ Prototype Feature Extraction mechanism (Sec. 3.1), the Cross-modal Correlation Studies between the Tissues and the Diseases (Sec. 3.2), and the Multi-modal Feature Fusion

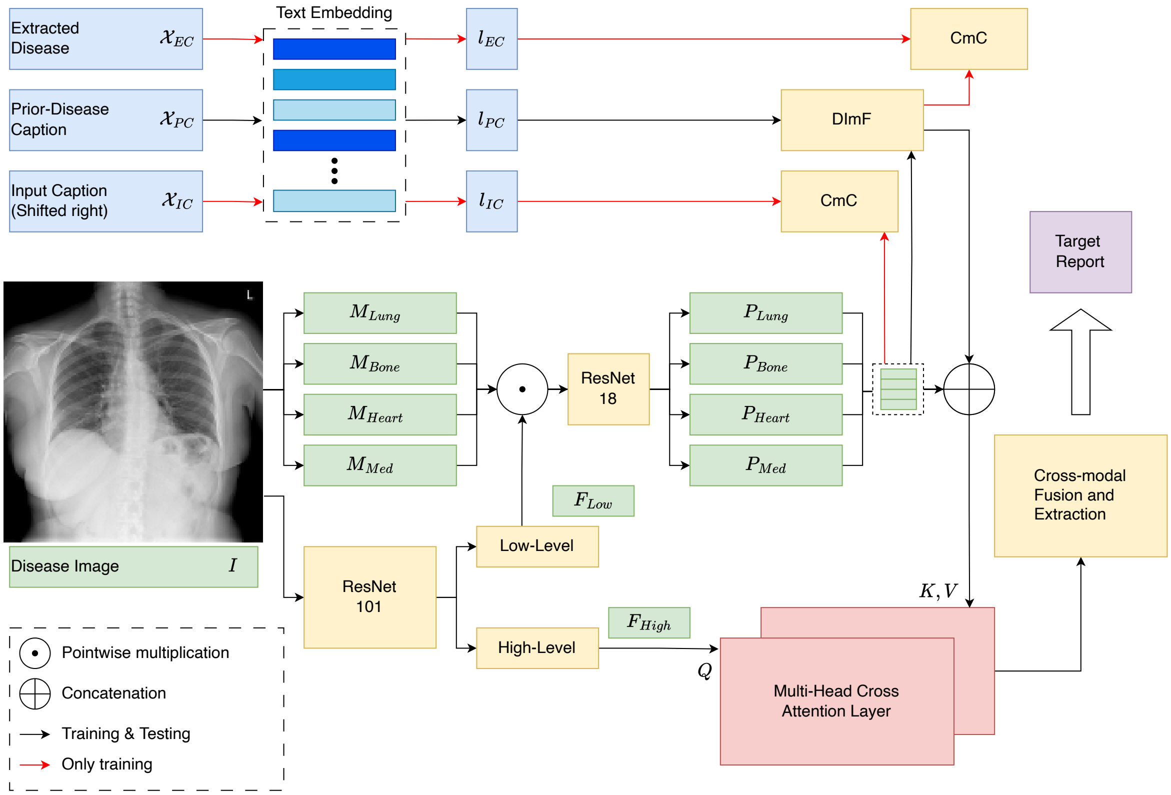


Figure 2. The overall architecture of our proposed COMG model. The *DImF* represents the fusion mechanisms between the embeddings for the prior-disease captions and the mask-guided prototype features, with details shown in Fig. 3, and *CmC* is the cross-modal consistency. Finally, the encoder-decoder structure generates the report. More detailed information is provided in Sec. 3.

and Consistency Mechanisms (Sec. 3.3).

3.1. Mask-guided Organ Prototype Feature Extraction

Due to most diseases existing on the body organs, the pixel-level mask information provided will help the model to particularly recognize these key areas [34]. To this end, we use the pre-trained CXAS model [32], which was trained on the PAX-RAY++ segmentation dataset [33]. By inferring the model on each X-ray image, it generates segmentation masks for various organs, including the heart, ribs, lung, and mediastinum (more specific information about masks is shown in the Supplementary Material). However, it's noteworthy that the COMG model is evaluated on two public benchmarks (IU-Xray, MIMIC-CXR), and the diseases mentioned in the reports are mainly related to four organs: bone, lung, heart, and mediastinum. Therefore, only partial masks belonging to these categories for each image will be employed. Since some organs are only related to limited types of disease, extracting the pixel-level masks for each organ and employing them for further analysis separately can induce the model to learn the correct correlations between the organs and disease, while ignoring negative pairing relationships.

After obtaining the masks for the key organs, we propose to extract the prototype features for each organ. The overall process is indicated in Fig. 2 and:

$$P_{og} = R_{ref}(F_{Low} \odot M_{og}), \tag{1}$$

where $og \in \{Bone, Lung, Heart, Mediastinum\}$. Specifically, the input images first pass through a ResNet 101 feature extractor for the intermediate features F_{Low}, and the final features F_{High}. Next, the resized F_{Low} is multiplied pointwise by the masks pre-extracted from each organ category (M_{og}). These multiplied features further pass through a ResNet18 (R_{ref}) for refinement to obtain the prototype features for the four key tissues. Based on the segmentation masks, these features contain semantic information for different key organs related to report generation (i.e., organ shapes and spatial relationships).

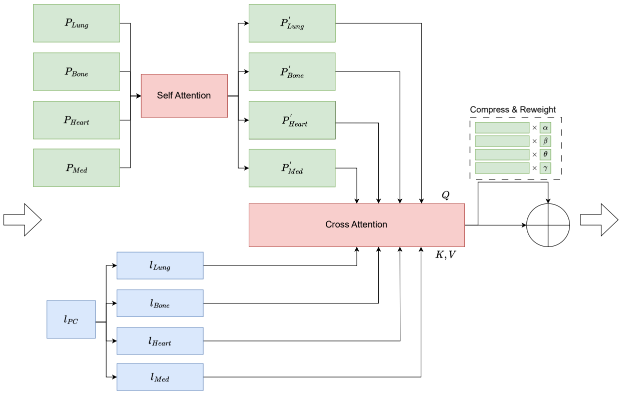


Figure 3. The structure of the disease image-mask fusion (*DImF*) module. The prototype features (P') of each organ will be inputted into a decoder structure with the corresponding disease token space (l) separately. **Reweight** are four learnable parameters to be multiplied with each corresponding organ's features. In addition, $\bigoplus$ means concatenation and **Compress** is the feature dimension compression. More specific information is provided in Sec. 3.2.

3.2. Cross-modal Correlation Studies between the Tissues and the Diseases

To accurately distinguish between diseases and normal cases, we combine the prototype features of each organ with the features from the keywords for the related diseases. The process is illustrated in Fig. 3. Referring to the disease symptom graph [14] (more details in the Supplementary Material), we can obtain the corresponding prior disease captions related to each organ. Note that these captions are based on the predefined knowledge graph and do not related to the report annotations of the images. Fig. 2 and Fig. 3, the correlation token for each organ (Tok_{og}) is calculated via:

$$Tok_{og} = CA\big(SA(P_{og}), l_{og}, l_{og}\big). \tag{2}$$

Specifically, the prototype features for each organ (P_{og} in Eq. 1) first pass through a self-attention SA layer. Next, the prior captions for each disease pass through transformer encoders for disease keyword embeddings. Within each organ type, the processed prototype features and the prior-disease caption embeddings l_{og} are fused via cross-attention mechanisms $CA(Q, K, V)$ to obtain the cross-modal correlation tokens Tok_{og} for the organ and its corresponding diseases. Specifically, the processed class-wise prototype features $SA(P_{og})$ are employed as the query, while the corresponding prior-disease caption embeddings l_{og} as the key and values. In addition, four (one for each tissue class) learnable parameters are developed to re-weight the importance of each cross-modal correlation token Tok_{og}. The tokens Tok_{og} from four organs are fused together according to these learnable parameters for a global correlation token Tok_{glb}.

3.3. Multi-modal Feature Fusion and Consistency

Multi-modal Feature Fusion. To facilitate the report generation learning via the comprehensive information aforementioned, we propose to fuse the global-level features directly extracted from the input images with the multi-modal features from the $DImF$ module. As indicated in Fig. 2, the output of the $DImF$ module Tok_{glb} is firstly reshaped and concated with each Organ Prototype Feature P_{og}. These concated features are employed as the key and value for the multi-head cross-attention layer, while the global-level image features F_{high} are the query. Such fused features contain knowledge from multiple perspectives, including the dense mask for key tissues, the text descriptions for the diseases, and the global-level features for the whole images. By passing them jointly with the embeddings for the extracted disease keywords from the report to transformer-based decoders, the quality of the generated report can be further improved.

Cross-modal Consistency. To further facilitate the feature extraction process under multiple modalities, we propose cross-modal consistency mechanisms (CmC module in Fig. 2). First, we propose to maximize the similarity between the embeddings of the input captions l_{IC} and the mask-guided organ prototype feature P_{og} (obtained from Sec. 3.1). According to the analysis of existing RRG methods [6, 35], the analysis in the reports is correlated to the specific regions and tissues in the radiology images. To this end, under the ideal feature extraction scenario, the organ prototype features should be dependent on the report descriptions. In addition, we also induce the similarity learning between the feature embeddings for the extracted disease for each image (l_{EC}), and the cross-modal correlation tokens (Tok_{og} from Sec. 3.2). For the cross-modal correlation tokens containing the text-level information regarding the diseases based on prior knowledge and the image-level information based on masks, their similarity with the disease keywords extracted from the ground truth should be maximized. It is because, under the optimal situation, the features for the multimodal tokens and the text keywords are both about the high-level characteristics of the diseases. For each similarity learning process, the two features are firstly resized into the same scale, then the cosine similarity loss (L_{sim}) [17] is utilized to enlarge their similarities. Specifically, the L_{sim} is defined as $L_{sim}(a, b) = 1 - \frac{a^T b}{||a||\ ||b||}$.

3.4. Training and Inference Details

The COMG model is optimized in two stages. The overall loss function $\mathcal{L}_{TI}$ for the first stage is three folds, defined

Datase	Methods	YEAR	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGEL
IU-Xray	R2Gen [6]	2020	0.470	0.304	0.219	0.165	0.187	0.371
	SEBTSAT+KG [47]	2020	0.441	0.291	0.203	0.147	-	0.304
	PPKED [25]	2021	0.483	0.315	0.224	0.168	0.190	0.376
	CMCL [24]	2022	0.473	0.305	0.217	0.162	0.186	0.378
	JPG [46]	2022	0.479	0.319	0.222	0.174	0.193	0.377
	CMM+RL [30]	2022	0.475	0.309	0.222	0.170	0.191	0.375
	KiUT [14]	2023	0.525	0.360	0.251	0.185	**0.242**	**0.409**
	METransformer [41]	2023	0.483	0.322	0.228	0.172	0.192	0.380
	DCL [21]	2023	-	-	-	0.163	0.193	0.383
	R2GenCMN*† [5]	2022	0.470	0.304	0.222	0.170	0.191	0.358
	COMG	**Ours**	0.482	0.316	0.233	0.184	0.198	0.382
	COMG + RL(Ours)	**Ours**	**0.536**	**0.378**	**0.275**	**0.206**	0.218	0.383
MIMIC-CXR	M2Transformer [8]	2020	0.332	0.210	0.142	0.101	0.134	0.264
	R2Gen [6]	2020	0.353	0.218	0.145	0.103	0.142	0.277
	PPKED [25]	2021	0.360	0.224	0.149	0.106	0.149	0.284
	CMCL [24]	2022	0.344	0.217	0.140	0.097	0.133	0.281
	CMM+RL [30]	2022	0.353	0.218	0.148	0.106	0.142	0.278
	UAR [22]	2023	0.363	0.229	0.158	0.107	0.157	0.289
	KiUT [14]	2023	**0.393**	**0.243**	0.159	0.113	**0.160**	0.285
	DCL [21]	2023	-	-	-	0.109	0.150	0.284
	R2GenCMN*† [5]	2022	0.348	0.206	0.135	0.094	0.136	0.266
	COMG	**Ours**	0.346	0.216	0.145	0.104	0.137	0.279
	COMG + RL(Ours)	**Ours**	0.363	0.235	**0.167**	**0.124**	0.128	**0.290**

Table 1. The results of the COMG model and other tested models in IU-Xray (upper part) and MIMIC-CXR (lower part) datasets. * indicates that we tested the results ourselves, which may differ from the results reported in the original papers of other models. † denotes the baseline model. The results for other models were obtained from their original papers. The best result is presented in bold, and the second-best result is underlined.

as:

$$\mathcal{L}_{TI} = \mathcal{L}_{CE} + \beta\mathcal{L}_{Sim_{IM}} + \theta\mathcal{L}_{Sim_{DT}}, \tag{3}$$

where CE is the cross-entropy loss for report generation study following [6], Sim_{IM} means the similarity-maximization loss between the embeddings of the input captions and the mask-guided organ prototype feature, and the Sim_{DT} is the cosine similarity loss between the cross-modal correlation tokens and extracted disease keywords features. We add tradeoff parameters to these loss functions to balance the overall optimization process, with β and θ set as 0.1. We have also presented the experimental analysis regarding different selections in the following sections.

After optimized via $\mathcal{L}_T$ for the first stage, we propose another stage of optimization for better performance by incorporating reinforcement learning (RL). Specifically, we included an additional BLEU score as a reward for RL to improve sentence coherence, combined with $\mathcal{L}_{CE}$ for report generation.

During inference, the cross-modal correlation tokens are first extracted from each image. Note that it is accessible since these tokens can be acquired by incorporating the image features with the masks from the pre-train segmentation models and the prior-disease captions from the pre-defined knowledge graph, which do not require ground-truth annotations. Then, the tokens are integrated with the image-level features via the multi-head attention. Finally, the decoder receives such fused features as input for report predictions.

4. Experiments

In this section, we first introduce the details of the experimental settings, including datasets, baseline models, and evaluation metrics. We then conduct the proposed COMG model on two datasets and evaluated it alongside some state-of-the-art approaches. In addition, ablation studies and the hyperparameters analysis of the COMG model are further presented.

4.1. Experiment Settings

4.1.1 Datasets

Two widely studied RRG benchmarks are employed to test the COMG model: IU-Xray [9] from Indiana University and MIMIC-CXR [15] from the Beth Israel Deaconess Medical Center. The MIMIC-CXR dataset is the largest publicly available radiography dataset, with 473,057 chest X-ray images and 206,563 associated reports. The IU-Xray is a relatively smaller dataset, which contains 7,470 chest X-ray images and 3,955 corresponding reports. Both datasets are divided into training, testing, and validation sets in a ratio of 7:2:1. More details of the datasets can be referred to the Supplementary Material. For both datasets, we followed Chen et al. [5] to pre-process captions and images. Before entering the model, each original radiology image was resized to 3 * 224 * 224 and normalized. In comparison, the mask images were resized to 1 * 224 * 224 to fit the mid-process fusion with mid-image features. To increase the model's robustness, images and masks were also enhanced with random cropping and random horizontal flipping. The captions were cleaned up, including removing punctuation and converting some words that appear less than three times to the token $< unk >$.

4.1.2 Baseline and Evaluation Metrics

Baseline. We compare the COMG model with nine existing radiology report generation models that have state-of-the-art (SOTA) results in the IU-Xray dataset. These models include R2Gen [6], PPKED [25], CMCL [24], KIUT [14], and METransformer [41]. We also compare the COMG model with the R2GenCMN [5] baseline model (marked † in Table 1).

In addition to the IU-Xray dataset, we compare the COMG model with nine SOTA models on a different dataset MIMIC-CXR, including M2Transformer [8], UAR [22], DCL [21], and the baseline model "R2GenCMN". The results of other comparison methods are cited from their respective papers, while the results of R2GenCMN are re-implemented by us as the baseline model (marked as * in Table 1).

Evaluation Metrics. We evaluate the quality of our report generation using natural language generation (NLG) metrics, *i.e.* BLEU [1-4] [29], METEOR [4] and ROUGE-L [23]. These metrics measure the similarity between the generated caption and the ground truth in terms of word-level n-grams. In addition, we follow the approach of [5, 6, 30] and use clinical efficacy (CE) metrics to evaluate the reports generated on the MIMIC-CXR dataset with their corresponding target captions. The CE metrics assess the presence of a set of significant clinical observations that can capture the diagnostic accuracy of the generated reports.

4.1.3 Implementation Details

We choose the ResNet101 pre-trained on ImageNet as the image extractor model and the CXAS [32] pre-trained on the radiology segmentation dataset PAX-RAY++ [33] as the mask extractor model. Our model is trained on a single NVIDIA GeForce RTX 3090 GPU with a 24GB memory. For optimization, following [5], we use the Adam optimizer [16]. The initial learning rates for the ResNet101 feature extractor and other components are set to 1e-4 and 5e-4, separately. During the inference stage, we incorporate the beam search [12] into the COMG model, with a step setting of 3. More experiment information has been provided in the Supplementary Material.

Method	CE Metric		
	Precision	Recall	F1
R2Gen [6]	0.333	0.273	0.276
CMM+RL [30]	0.342	0.294	0.292
METransformer [41]	0.364	0.309	0.311
KiUT [14]	0.371	**0.318**	0.321
R2GenCMN*† [5]	0.334	0.275	0.278
COMG(Ours)	**0.424**	0.291	**0.345**

Table 2. Comparison of clinical efficacy metrics for the MIMIC-CXR dataset. These metrics measure the accuracy of clinical abnormality descriptions. The best result is presented in bold, and the second-best result is underlined.

4.2. Experiment Results and Analysis

4.2.1 Radiology Report Generation

Two evaluation metrics are used for comparison: conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics. These are common metrics used to evaluate the report generation task. The results are shown in Table 1 and Table 2, respectively.

Descriptive Accuracy. We report the descriptive accuracy in Table 1. As can be seen from the results on IU-Xray, the COMG model outperforms the baseline model "R2GenCMN" in all aspects, including BLEU [1,2,3,4], Meteor, and Rouge-L by adding all the contributions mentioned in Sec. 1. By adding reinforcement learning as a second step in training for the COMG model, our model excels in BLEU [1,2,3,4] and achieves the second best results in METEOR and ROUGE-L. In radiology report generation, BLEU@4 is an important guideline [5], and the COMG model achieves a significant improvement in this metric compared to "KiUT" (*i.e.*, 0.185 → 0.206).

In the MIMIC-CXR dataset, the "RGRG" and "METransformer" are excluded. The "RGRG" used a very large model with a 24-layer decoder, making it difficult for others to reproduce its results. The "METransformer" did not make its code public, which makes it impossible to rerun the experiment on the MIMIC-CXR dataset. Furthermore, the dataset split used in our evaluation was different from these two models. Compared to the baseline model "R2GenCMN", our model has significant improvements in all six evaluation metrics. By adding RL to the COMG model, our model achieves substantial performance gains in BLEU [3,4] and ROUGE-L, and achieves the second-

#	Method	IU-Xray							MIMIC-CXR						
		B@1	B@2	B@3	B@4	MET.	RGL.	AVG.	B@1	B@2	B@3	B@4	MET.	RGL.	AVG.
1	Baseline	0.462	0.299	0.221	0.172	0.193	0.37	-	0.330	0.201	0.134	0.094	0.134	0.269	-
2	+ Mk	0.504	0.323	0.227	0.170	0.192	0.388	↓ 1.1%	0.338	0.206	0.138	0.098	0.130	0.270	↑ 4.3%
3	+ Mk + Sim$_{IM}$	0.469	0.303	0.223	0.175	0.187	0.367	↑ 1.7%	0.343	0.211	0.141	0.100	0.134	0.275	↑ 6.4%
4	+ Mk + DT + Sim$_{DT}$	0.484	0.320	0.234	0.182	0.204	0.379	↑ 5.8%	0.347	0.213	0.142	0.102	0.136	0.275	↑ 8.5%
5	+ Mk + Sim$_{IM}$ + DT + Sim$_{DT}$	0.482	0.316	0.233	0.184	0.198	0.382	↑ 7.0%	0.347	0.216	0.145	0.104	0.137	0.279	↑ 10.6%
6	+ Mk + Sim$_{IM}$ + DT + Sim$_{DT}$ + RL	0.536	0.378	0.275	0.206	0.218	0.383	↑ 20.0%	0.363	0.235	0.167	0.124	0.128	0.290	↑ 31.9%

Table 3. The ablation study of the COMG model on the IU-Xray and MIMIC-CXR datasets. AVG indicates the improvement in the BLEU@4 value compared to the baseline model, while RL stands for reinforcement learning. MET. and RGL. represent Meteor and Rouge-L, respectively.

best result in BLEU [1,2] compared to other existing models, *e.g*., the BLEU@4 score increased by 9.7% compared to the second-best result from "KiUT". However, the result of METEOR has dropped slightly when RL is added, because the reward in RL is based on BLEU metrics while ignoring the METEOR score. We have also tried to employ other metrics as rewards for RL, and we noticed their performance is inferior to the RL with BLEU.

Clinical Correctness. Table 2 reports the quantitative results of our proposed model and 4 SOTAs, *i.e*., R2Gen [6], MEtransformer [41], KiUT [14], and the baseline model R2GenCMN [5], on the MIMIC-CXR dataset. As can be seen, our COMG model performs better in precision and F1 score than all SOTAs, *e.g*., precision increased by 14.3% and F1 increased by 7.5%. Compared to "R2GenCMN", COMG achieves a significant improvement on all metrics.

4.2.2 Ablation Study

In this section, we conduct an ablation study to investigate the effect of each designed module in our approach. Table 3 shows the experimental results on two datasets: IU-Xray and MIMIC-CXR. Specifically, MK represents the introduction of the mask-guided organ prototype features for model training (Sec. 3.1). Sim_{IM} represents including the similarity loss $\mathcal{L}_{Sim_{IM}}$, while Sim_{DT} represents adding the similarity loss $\mathcal{L}_{Sim_{DT}}$. DT represents adding the cross-modal correlation token (Sec. 3.2), and RL represents the addition of reinforcement learning. We evaluated the model using three kinds of metrics: BLEU [1,2,3,4], METEOR, and ROUGE-L. The AVG. column shows the average increase of each model compared to the baseline model, based on the BLEU@4 metric, which is the most important evaluation metric in radiology report generation tasks.

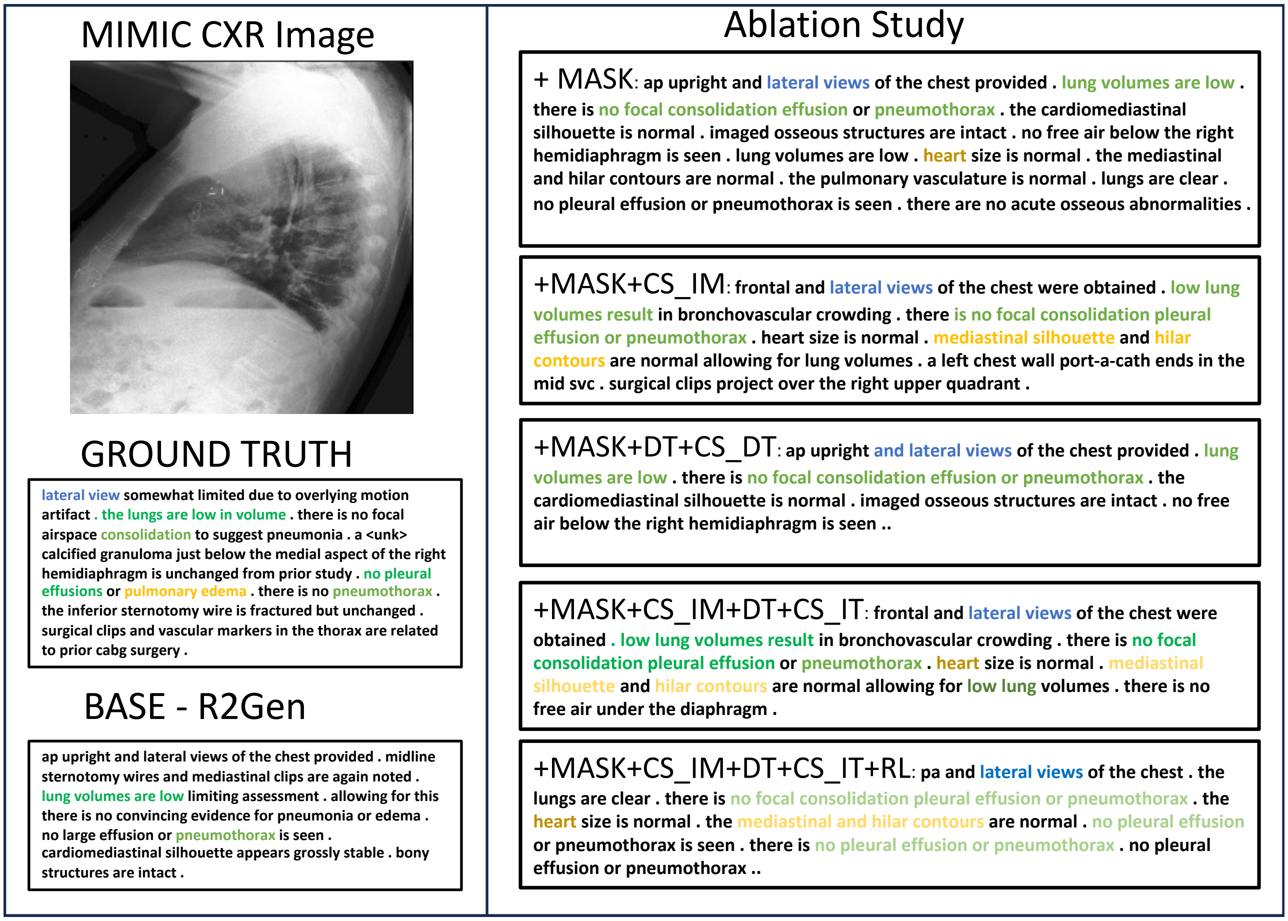
By comparing #1 and #2 in Table 3, it indicates that incorporating the mask-guided organ prototype features has made the greatest contribution to the improvement of the model's performance, which indicates the importance of the pixel-level information for report generation. Additionally, we have also validated the effectiveness of the two cross-modal consistency mechanisms. The performance gain for #3 over #2 shows the benefits of enlarging the cross-modal similarity between the organ prototype features and the report descriptions. By jointly employing the cross-modal correlation tokens with its similarity learning with the disease keywords from ground truth, the baseline has been improved by a large margin (#4). Next, the improvement of the first-stage model #5 over the baseline #1 has further indicated the model's effectiveness by jointly integrating all proposed modules. Finally, #6 shows the result of combining all contributions and adding reinforcement learning. It significantly improves performance in most metrics, while only losing fewer marks of Rouge-L in IU-Xray and Meteor in MIMIC-CXR due to the rewards being set on the BLEU metrics.

4.2.3 Qualitative Analysis

To further investigate the effectiveness of our method, we perform qualitative analysis on the MIMIC-CXR dataset (shown in Fig. 4). In the example, we have highlighted the keywords related to organs and diseases in distinct colors for clear differentiation. It shows that if the model detects certain parts of the disease incorrectly, its prediction will fail to generate the corresponding descriptions successfully. More specifically, baseline can only detect lung and pneumothorax, while our COMG can detect more details about mediastinal and hilar contours, which makes our report more vivid and accurate. More examples can be found in the Supplementary Material.

4.2.4 Hyper-parameter Analysis

Table 4 shows the results of different combinations of loss function coefficients β, and θ. We changed the coefficients of β and θ from 0.1 to 10 to evaluate the influence of each



MIMIC CXR Image

GROUND TRUTH

lateral view somewhat limited due to overlying motion artifact . the lungs are low in volume . there is no focal airspace consolidation to suggest pneumonia . a <unk> calcified granuloma just below the medial aspect of the right hemidiaphragm is unchanged from prior study . no pleural effusions or pulmonary edema . there is no pneumothorax . the inferior sternotomy wire is fractured but unchanged . surgical clips and vascular markers in the thorax are related to prior cabg surgery .

BASE - R2Gen

ap upright and lateral views of the chest provided . midline sternotomy wires and mediastinal clips are again noted . lung volumes are low limiting assessment . allowing for this there is no convincing evidence for pneumonia or edema . no large effusion or pneumothorax is seen . cardiomediastinal silhouette appears grossly stable . bony structures are intact .

Ablation Study

+ MASK: ap upright and lateral views of the chest provided . lung volumes are low . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen . lung volumes are low . heart size is normal . the mediastinal and hilar contours are normal . the pulmonary vasculature is normal . lungs are clear . no pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities .

+MASK+CS_IM: frontal and lateral views of the chest were obtained . low lung volumes result in bronchovascular crowding . there is no focal consolidation pleural effusion or pneumothorax . heart size is normal . mediastinal silhouette and hilar contours are normal allowing for lung volumes . a left chest wall port-a-cath ends in the mid svc . surgical clips project over the right upper quadrant .

+MASK+DT+CS_DT: ap upright and lateral views of the chest provided . lung volumes are low . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal . imaged osseous structures are intact . no free air below the right hemidiaphragm is seen ..

+MASK+CS_IM+DT+CS_IT: frontal and lateral views of the chest were obtained . low lung volumes result in bronchovascular crowding . there is no focal consolidation pleural effusion or pneumothorax . heart size is normal . mediastinal silhouette and hilar contours are normal allowing for low lung volumes . there is no free air under the diaphragm .

+MASK+CS_IM+DT+CS_IT+RL: pa and lateral views of the chest . the lungs are clear . there is no focal consolidation pleural effusion or pneumothorax . the heart size is normal . the mediastinal and hilar contours are normal . no pleural effusion or pneumothorax is seen . there is no pleural effusion or pneumothorax . no pleural effusion or pneumothorax ..

Figure 4. An example of the report generated by different models in the ablation study. The left side of the image displays the input image from the MIMIC-CXR dataset, the corresponding ground truth, and the report generated by the baseline model. In the image, we have marked the keywords of organs and diseases in different colors other than black. More specific information is shown in Sec. 4.2.3.

	Loss Coefficient						
β	0.1	0.1	1	1	10	1	10
θ	0.1	1	0.1	1	1	10	10
B@1	**0.482**	0.444	0.433	0.470	0.468	0.435	0.418
B@2	**0.316**	0.282	0.273	0.306	0.300	0.178	0.274
B@3	**0.233**	0.200	0.190	0.213	0.215	0.198	0.197
B@4	**0.184**	0.148	0.142	0.158	0.162	0.150	0.148
MET.	**0.198**	0.174	0.177	0.197	0.190	0.178	0.174
RGL.	**0.382**	0.341	0.342	0.366	0.365	0.360	0.360

Table 4. The influence of each coefficient on each loss component in the loss function. This experiment was conducted using the IU-Xray dataset.

loss function coefficient. Table 4 lists the results of our COMG under this range. We find that 0.1, and 0.1 are the best choices for β, and θ, respectively.

5. Conclusion

In this paper, we propose a novel COMG method for generating precise radiology reports. It employs complex organ masks to provide pixel-wise semantic information for accurate report generation. Additionally, it incorporates disease keywords linked to each tissue, utilizing them as text-level prior knowledge to further refine tissue feature learning. To streamline the feature extraction process for both images and text, we have developed two cross-modal consistency mechanisms to enhance feature learning accuracy. Our method has been tested on two popular benchmarks, and the results show its effectiveness in generating accurate and meaningful reports. In future works, we plan to enhance the COMG's ability to recognize abnormal tissues/regions by incorporating external resources (e.g., Chest ImaGenome dataset [42]). This can further improve the reports' quality on disease recognition and understanding.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021. 2

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[4] Satanjeev Banerjee and Alon Lavie. 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, 2005. 6

[5] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Aug. 2021. 1, 2, 5, 6, 7

[6] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Nov. 2020. 1, 2, 4, 5, 6, 7

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 1, 2

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 5, 6

[9] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 5

[10] Shubham Dodia, B Annappa, and Padukudru A Mahesh. Recent advancements in deep learning based lung cancer detection: A systematic review. *Engineering Applications of Artificial Intelligence*, 116:105490, 2022. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017. 6

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[14] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023. 4, 5, 6, 7

[15] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6, 2016. 4

[18] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2

[19] Haoran Li, Chun-Mei Feng, Yong Xu, Tao Zhou, Lina Yao, and Xiaojun Chang. Zero-shot camouflaged object detection. *IEEE Transactions on Image Processing*, pages 5126–5137, 2023. 2

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[21] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. 1, 2, 5, 6

[22] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. Unify, align and refine: Multi-level semantic alignment for radiology report generation. *arXiv e-prints*, pages arXiv–2303, 2023. 5, 6

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[24] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3001–3012, Online, Aug. 2021. Association for Computational Linguistics. 5, 6

[25] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 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, 2021. 5, 6

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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*, pages 375–383, 2017. 1, 2

[27] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 2

[28] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 2

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 6

[30] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022. 1, 2, 5, 6

[31] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 1

[32] Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. *arXiv preprint arXiv:2306.03934*, 2023. 1, 2, 3, 6

[33] Constantin Marc Seibold, Simon Reiß, M. Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 3, 6

[34] Paul Suetens, Erwin Bellon, Dirk Vandermeulen, M Smet, Guy Marchal, Johan Nuyts, and Luc Mortelmans. Image segmentation: methods and applications in diagnostic radiology and nuclear medicine. *European journal of radiology*, 17(1):14–21, 1993. 3

[35] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 1, 2, 4

[36] Ajay K Tanwani, Joelle Barral, and Daniel Freedman. Repsnet: Combining vision with language for automated medical reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 714–724. Springer, 2022. 1, 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1

[39] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer, 2022. 1, 2

[40] Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012. 1

[41] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 2, 5, 6, 7

[42] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. 8

[43] Ting-Wei Wu, Jia-Hong Huang, Joseph Lin, and Marcel Worring. Expert-defined keywords improve interpretability of retinal image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2023. 2

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[45] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 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, 2021. 2

[46] Jingyi You, Dongyuan Li, Manabu Okumura, and Kenji Suzuki. Jpg-jointly learn to align: Automated disease prediction and radiology report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5989–6001, 2022. 5

[47] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917, 2020. 5