Interactive and Explainable Region-guided Radiology Report Generation

Tim Tanida[1,*] Philip Müller[1,*] Georgios Kaissis[1,2] Daniel Rueckert[1,3]
[1]Technical University of Munich, [2]Helmholtz Zentrum Munich, [3]Imperial College London
{tim.tanida, philip.j.mueller, g.kaissis, daniel.rueckert}@tum.de

Abstract

The automatic generation of radiology reports has the potential to assist radiologists in the time-consuming task of report writing. Existing methods generate the full report from image-level features, failing to explicitly focus on anatomical regions in the image. We propose a simple yet effective region-guided report generation model that detects anatomical regions and then describes individual, salient regions to form the final report. While previous methods generate reports without the possibility of human intervention and with limited explainability, our method opens up novel clinical use cases through additional interactive capabilities and introduces a high degree of transparency and explainability. Comprehensive experiments demonstrate our method's effectiveness in report generation, outperforming previous state-of-the-art models, and highlight its interactive capabilities. The code and checkpoints are available at https://github.com/ttanida/rgrg.

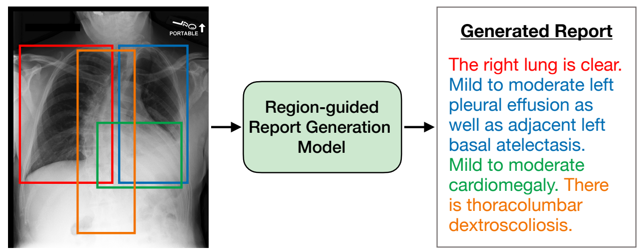

Figure 1. Our approach at a glance. Unique anatomical regions of the chest are detected, the most salient regions are selected for the report and individual sentences are generated for each region. Consequently, each sentence in the generated report is explicitly grounded on an anatomical region.

1. Introduction

Chest radiography (chest X-ray) is the most common type of medical image examination in the world and is critical for identifying common thoracic diseases such as pneumonia and lung cancer [19, 37]. Given a chest X-ray, radiologists examine each depicted anatomical region and describe findings of both normal and abnormal salient regions in a textual report [12]. Given the large volume of chest X-rays to be examined in daily clinical practice, this often becomes a time-consuming and difficult task, which is further exacerbated by a shortage of trained radiologists in many healthcare systems [3, 40, 41]. As a result, automatic radiology report generation has emerged as an active research area with the potential to alleviate radiologists' workload.

Generating radiology reports is a difficult task since reports consist of multiple sentences, each describing a specific medical observation of a specific anatomical region. As such, current radiology report generation methods tend to generate reports that are factually incomplete (*i.e.*, missing key observations in the image) and inconsistent (*i.e.*, containing factually wrong information) [28]. This is further exacerbated by current methods utilizing image-level visual features to generate reports, failing to explicitly focus on salient anatomical regions in the image. Another issue regarding existing methods is the lack of explainability. A highly accurate yet opaque report generation system may not achieve adoption in the safety-critical medical domain if the rationale behind a generated report is not transparent and explainable [11, 14, 27]. Lastly, current methods lack interactivity and adaptability to radiologists' preferences. *E.g.*, a radiologist may want a model to focus exclusively on specific anatomical regions within an image.

Inspired by radiologists' working patterns, we propose a simple yet effective **R**egion-**G**uided Radiology **R**eport **G**eneration (RGRG) method to address the challenges highlighted before. Instead of relying on image-level visual features, our work is the first in using object detection to directly extract localized visual features of anatomical regions, which are used to generate individual, anatomy-specific sentences describing any pathologies to form the final report. Conceptually, we *divide and conquer* the difficult task of generating a complete and consistent report from the whole image into a series of simple tasks of gen-

*Equal contribution

erating short, consistent sentences for a range of isolated anatomical regions, thereby achieving both completeness and consistency for the final, composed report.

While existing models [7, 17, 26] can produce heatmaps to illustrate which parts of an image were attended during report generation, our model visually grounds each sentence in the generated report on a predicted bounding box around an anatomical region (see Fig. 1), thereby introducing a high degree of explainability to the report generation process. The visual grounding allows radiologists to easily verify the correctness of each generated sentence, which may increase trust in the model's predictions [11, 27, 39]. Moreover, our approach enables interactive use, allowing radiologists to select anatomical structures or draw bounding boxes around regions of interest for targeted description generation, enhancing flexibility in the clinical workflow.

Our contributions are as follows:

- We introduce a simple yet effective **R**egion-**G**uided Radiology **R**eport **G**eneration (RGRG) method that detects anatomical regions and generates individual descriptions for each. Empirical evidence demonstrates that our model produces relevant sentences pertaining to the anatomy. To the best of our knowledge, it is the first report generation model to visually ground sentences to anatomical structures.
- We condition a pre-trained language model on each anatomical region independently. This enables **anatomy-based sentence generation**, where radiologists interactively select individual, detected anatomical regions for which descriptions are generated.
- Additionally, our approach enables radiologists to manually define regions of interest using bounding boxes, generating corresponding descriptions. We assess the impact of these manually drawn boxes, demonstrating the robustness of the **selection-based sentence generation** task.
- For **full radiology report generation**, a module is introduced to select salient anatomical regions, whose generated descriptions are concatenated to form factually complete and consistent reports. We empirically demonstrate this on the MIMIC-CXR dataset [18, 19], where we outperform competitive baselines in both language generation and clinically relevant metrics.

2. Related work

2.1. Radiology report generation

Transformer-based radiology report generation. Automatic radiology report generation has garnered significant research interest in recent years. While early works [16, 17, 22, 48, 56] adopted CNN-RNN architectures, more recent studies have shifted towards utilizing the effectiveness of the Transformer [44]. The standard transformer architecture was adapted by adding relation memory units [7] or a memory matrix [6] to improve cross-model interactions between visual and textual features. To better attend to abnormal regions, a transformer was proposed that iteratively aligns extracted visual features and disease tags [53], a system contrasting normal and abnormal images was presented [25], and an additional medical knowledge graph was utilized [24]. In contrast, we focus on abnormal regions by directly extracting the corresponding region visual features via object detection, additionally encoding strong abnormal information in the features with an abnormality classification module. Recently, warm-starting components from pre-trained models has shown promising results for report generation [29]. Inspired by [1], our work uses a pre-trained language model as the decoder.

Consecutive radiology report generation. Similar to our approach, a variety of previous works have attempted to divide the difficult task of generating a long, coherent radiology report into several steps. Hierarchical approaches were proposed [26, 30], in which high-level concepts are first extracted from the image and subsequently decoded into individual sentences. Our work is most related to [47]. They introduce a multi-head transformer applied to patch features from a CNN backbone, with each head assigned to a specific anatomical region, generating sentences exclusively for that region. In contrast, our method extracts region-level features through object detection (as opposed to image-level features), and generates region-level sentences with a shared transformer decoder based on these features (instead of selectively using designated heads). While our method requires more supervision, it is overall less technically complex, whilst offering additional interactivity and a higher degree of transparency and explainability.

2.2. Image captioning

Image captioning. Most radiology report generation methods [1,6,7,17,24,25,48,53] are inspired by influential works [9, 46, 51, 54] from the image captioning domain in computer vision. While overarching concepts can be translated from general image captioning to radiology report generation, there are notable differences: 1) Radiology reports are much longer and more diverse than typical captions, describing multiple anatomical regions in an image. 2) Generating descriptions of specific but crucial abnormalities is complicated because of a heavy data bias towards normal images and normal reports.

Dense image captioning. Our work is related to image captioning methods that utilize object detection in the encoding process, in particular works [20, 21, 43, 52] from the dense image captioning domain. Instead of generating a single caption for the whole image, dense captioning methods aim to both localize and describe individual salient regions in images, usually by conditioning a language model on the specific region features. Intuitively, dense image captioning aligns more closely with radiologists' working practice, as anatomical regions in an image are usually localized and described one by one when crafting a report.

Controllable image captioning. In our anatomy-based sentence generation mode (see Sec. 1), radiologists can manually select anatomical regions automatically outlined by bounding boxes, which are then individually described by generated sentences. This is related to the domain of controllable image captioning [5, 8, 57], in which generated captions are influenced by an external control signal, *e.g.* by specifying the image regions to be described [8, 57], or by specifying an abstract scene graph [5]. However, while controllable image captioning methods typically generate a single image-level caption, our model in this mode generates descriptions independently for each selected region.

3. Method

Our proposed **R**egion-**G**uided Radiology **R**eport **G**eneration (RGRG) method has three distinct use cases:

1. **Radiology report generation**: The proposed model generates a full radiology report for a chest X-ray image and outlines the described anatomical regions by bounding boxes in the image, as depicted in Fig. 1.

2. **Anatomy-based sentence generation**: The proposed model outlines anatomical regions by bounding boxes. The radiologist then selects individual anatomies of interest, for which the model generates sentences.

3. **Selection-based sentence generation**: The radiologist manually draws a bounding box around a region of interest, for which the model then generates sentences.

We first outline our model in the context of radiology report generation and elaborate on anatomy- and selection-based sentence generation in Sec. 3.4.

3.1. Overview

For full report generation, our model closely follows the typical workflow of a radiologist. First, a radiologist identifies distinct anatomical regions in an X-ray image. For each region, a decision is made if the region needs to be described in the report and if it is abnormal, with the latter assessment influencing the former. Finally, each selected region is described in 1-2 short sentences in the final report.

Fig. 2 illustrates the proposed model architecture consisting of four major modules. First, an object detector identifies and extracts visual features of 29 distinct anatomical regions of the chest. Radiologists usually only describe a handful of salient regions in a report (see the exemplary report of Fig. 1). The region selection module simulates this behavior by predicting via a binary classifier whether sentences should be generated for each region, *e.g.*, the image of Fig. 2 has four regions selected for sentence generation.

During training, the region visual features are also fed into the abnormality classification module consisting of an additional binary classifier that predicts whether a region is normal or abnormal (*i.e.*, contains a pathology). This encodes stronger abnormality information in the region visual features from the object detector, helping both the region selection module in selecting abnormal regions for sentence generation, as well as the decoder in generating sentences that capture potential pathologies of the regions.

The decoder is a transformer-based language model pre-trained on medical abstracts. It generates sentences for each selected region (treated as independent samples) by conditioning on the associated region visual features. The final report is obtained via a post-processing step which removes generated sentences that are too similar to each other and concatenates the remaining ones.

3.2. Modules

Object detector. For the object detector, we use Faster R-CNN [38] with a ResNet-50 [15] backbone pre-trained on ImageNet [10]. Faster R-CNN consists of a region proposal network (RPN), which generates object proposals (*i.e.*, bounding boxes of potential anatomical regions) based on the feature maps extracted by the backbone from the input image. A region of interest (RoI) pooling layer maps each object proposal onto the backbone feature maps, extracting small feature maps of uniform spatial extent for the proposals. These RoI feature maps are each classified into one of the 29 anatomical region classes (and the background class) following standard procedure in Faster R-CNN.

To obtain the region visual features for each of the 29 anatomical regions, we first determine the "top" object proposal for each region class. A "top" object proposal for a given region class is one where the class has the highest probability score (of all classes) for the given proposal, as well as the highest score for all proposals where the same region class is also the top-1 class. If a region class does not achieve the highest score for at least one proposal, it is considered an undetected class. Undetected classes are automatically deselected by the region selection module.

The region visual features $\in \mathbb{R}^{29 \times 1024}$ are then obtained by taking the 29 RoI pooling layer features maps $\in \mathbb{R}^{29 \times 2048 \times H \times W}$ corresponding to the 29 "top" object proposals, applying 2D average pooling over the spatial dimensions and reducing the dimension from 2048 to 1024 by linear transformation.

Region selection and abnormality classification. The binary classifiers of these modules consist of multilayer per-

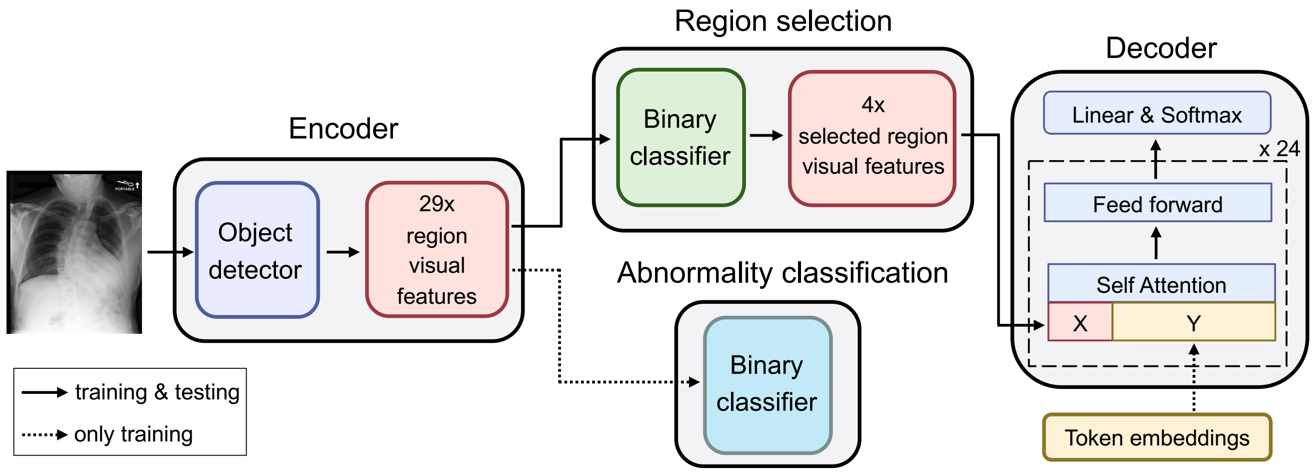

Figure 2. **R**egion-**G**uided Radiology **R**eport **G**eneration (RGRG): the object detector extracts visual features for 29 unique anatomical regions of the chest. Two subsequent binary classifiers select salient region features for the final report and encode strong abnormal information in the features, respectively. The language model generates sentences for each of the selected regions (in this example 4), forming the final report. For conciseness, residual connections and layer normalizations in the language model are not depicted.

ceptrons that gradually reduce the input features to a single output logit to predict if a region is selected for sentence generation and is abnormal, respectively.

Language model. For the language model, we use the 355M-parameter model GPT-2 Medium [36] fine-tuned on PubMed abstracts [31]. GPT-2 is an auto-regressive neural network based on self-attention, in which tokens in a sequence are conditioned on previous tokens for text generation. This can be expressed as (ignoring scaling factors):

$$\mathrm{SA}(Y) = \mathrm{softmax}((YW_q)(YW_k)^\top)(YW_v), \qquad (1)$$

where Y represents the token embeddings, and W_q, W_k, W_v are the query, key, and value projection parameters.

To condition the language model on the region visual features, we follow [1] in using pseudo self-attention [58] to directly inject the region visual features in the self-attention of the model, *i.e*.

$$\mathrm{PSA}(X, Y) = \mathrm{softmax}\left((YW_q)\begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^\top\right)\begin{bmatrix} XU_v \\ YW_v \end{bmatrix}, \qquad (2)$$

where X represents the region visual features, and U_k and U_v are the corresponding (newly initialized) key and value projection parameters. This allows text generation conditioned on both previous tokens and region visual features.

3.3. Training

The model training consists of three steps. First, only the object detector is trained, then the object detector combined with the binary classifiers, and finally the full model end-to-end with all parameters left trainable.[1] To train the language model, we only use region visual features that have corresponding reference sentences, assuming that the region selection module correctly selects these regions (that require sentences) at test time. For regions with several sentences, the sentences are concatenated, such that the model learns to predict several sentences in such cases.

The overall training loss is defined as

$$\begin{aligned} \mathcal{L} = {} & \lambda_{\text{obj}} \cdot \mathcal{L}_{\text{obj}} + \lambda_{\text{select}} \cdot \mathcal{L}_{\text{select}} \\ & + \lambda_{\text{abnormal}} \cdot \mathcal{L}_{\text{abnormal}} + \lambda_{\text{language}} \cdot \mathcal{L}_{\text{language}}\,, \end{aligned} \qquad (3)$$

where $\mathcal{L}_{\text{obj}}$ is the Faster R-CNN object detector loss, $\mathcal{L}_{\text{select}}$ and $\mathcal{L}_{\text{abnormal}}$ are the (weighted) binary cross-entropy losses for the two binary classifiers, and $\mathcal{L}_{\text{language}}$ is the cross-entropy loss for the language model. Based on the performance on the validation set, the loss weights are set as λ_{obj} = 1.0, λ_{select}= 5.0, $\lambda_{\text{abnormal}}$= 5.0, and $\lambda_{\text{language}}$= 2.0.

3.4. Inference

Radiology report generation. Reports are formed by concatenating the generated sentences of selected anatomical regions. If pathologies span several such regions or multiple anatomically similar regions (*e.g*., left and right lung) have no findings, then similar sentences may be generated for these regions, leading to duplicate sentences in the generated report. To remove such duplicates, we use BERTScore [55] to determine the degree of similarity, always removing the shorter sentence and keeping the longer one, as longer sentences tend to contain more clinically relevant information. *E.g*., the sentences *"The cardiomediastinal silhouette and hilar contours are normal."* and *"The cardiomediastinal silhouette is normal."* are quite similar, while the longer sentence contains additional, relevant information.

[1] Note that following [58], in the language model only the projection parameters tasked with injecting the region visual features are trainable.

Anatomy-based sentence generation. In this mode, radiologists can manually select anatomical regions (from the 29 overall regions) for the model to examine. First, the model detects and outlines all 29 regions by bounding boxes via the object detector. Next, the radiologists select regions of interest, which are in turn (exclusively) selected by the region selection module for sentence generation.

Selection-based sentence generation. In this mode, radiologists can draw bounding boxes around individual anatomical regions for examination. The bounding boxes are passed through the RoI pooling layer and further transformed into region visual features (as described in Sec. 3.2), which are then fed into the language model for sentence generation.

4. Experimental setup

We evaluate our method on all three tasks: **radiology report generation**, **anatomy-based sentence generation**, and **selection-based sentence generation** (see Sec. 3). We refer to the supplementary materials for more details on the experimental setup.

4.1. Dataset and pre-processing

We use the Chest ImaGenome v1.0.0 [13, 49, 50] dataset to train and evaluate our proposed model. It is constructed from the MIMIC-CXR [18, 19] dataset, which consists of chest X-ray images with corresponding free-text radiology reports. The Chest ImaGenome dataset contains automatically constructed scene graphs for the MIMIC-CXR images. Each scene graph describes one frontal chest X-ray image and contains bounding box coordinates for 29 unique anatomical regions in the chest, as well as sentences describing each region if they exist in the corresponding radiology report. We use the official split provided by the dataset resulting in 166,512 training images, 23,952 validation images, and 47,389 test images.

All images are resized to 512x512 while preserving the original aspect ratio, padded if needed, and normalized to zero mean and unit standard deviation. Color jitter, Gaussian noise, and affine transformations are applied as image data augmentations during training. For the sentences, we remove redundant whitespaces (*i.e*., line breaks, etc.). For the radiology report generation task, we follow previous work [4, 26, 28, 29] in using the *findings* section of the radiology reports (of the MIMIC-CXR dataset) as our reference reports. The *findings* section contains the observations made by the radiologist. Following [28, 29], we discard reports with empty *findings* sections, resulting in 32,711 test images with corresponding reference reports. We note that no further processing is applied to the extracted reports (as opposed to *e.g*. [29]).

4.2. Evaluation metrics

On report level, we evaluate the model on widely used natural language generation (NLG) metrics such as BLEU [32], METEOR [2], ROUGE-L [23] and CIDEr-D [45], which measure the similarity between generated and reference report by counting matching n-grams (*i.e*., word overlap). On sentence level, we evaluate on METEOR, a metric appropriate for evaluation on both sentence- and report-level (as opposed to *e.g*. BLEU). Since conventional NLG metrics are ill-suited to measure the clinical correctness of generated reports [4,26,34], we follow [7,26,28,29] in additionally reporting clinical efficacy (CE) metrics. CE metrics compare generated and reference reports w.r.t. the presence status of an array of prominent clinical observations, thus capturing the diagnostic accuracy of generated reports.

4.3. Evaluation strategy and baselines

Radiology report generation. We compare our model with previous state-of-the-art models, specifically R2Gen [7], CMN [6], PPKED [24], $\mathcal{M}^2$ TR. PROGRESSIVE [30], Contrastive Attention [25], AlignTransformer [53], $\mathcal{M}^2$ Trans [9] optimized on a standard language model loss and rewards towards factual completeness and consistency [28], ITA [47], CvT-212DistilGPT2 [29]. If not denoted otherwise, we cite the results from the corresponding papers.

Anatomy-based sentence generation. To the best of our knowledge, this is the first work in using a model capable of generating sentences for specific anatomical regions of a chest X-ray. As such, we have no direct baselines to compare against and therefore conduct a qualitative analysis. Additionally, we report the performance of the language model in this task. Specifically, we report the per-anatomy METEOR scores (*i.e*., computed between the generated and reference sentences of each anatomy independently) for six prominent regions as well as the micro-average over all, normal, and abnormal regions, respectively. Additionally, for further verification of the model generating anatomy-sensitive sentences, we compute a custom "Anatomy-Sensitivity-Ratio" (short AS-Ratio). The ratio is computed by dividing the micro-averaged METEOR score over all regions by an "anatomy-agnostic" METEOR score, calculated when generated sentences of an anatomy are (falsely) paired with reference sentences of all other anatomies of an image. *E.g*., if hypothetically there were only the 3 anatomical regions of *right lung* (RL), *left lung* (LL) and *spine* (SP), then the "anatomy-agnostic" score would be calculated by pairing the generated sentences of RL with the reference of LL and SP, the generated sentences of LL with the reference of RL and SP etc., for a given image. Consequently, the AS-Ratio measures how closely aligned the generated sentences are to their corresponding anatomy, with a ratio of 1.0 representing no apparent relation, while higher ratios represent closer alignment.

Dataset	Method	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MIMIC-CXR	R2Gen [7]	2020	0.353	0.218	0.145	0.103	0.142	0.277	0.406†
	CMN [6]	2021	0.353	0.218	0.148	0.106	0.142	0.278	-
	PPKED [24]	2021	0.360	0.224	0.149	0.106	0.149	0.284	0.237
	$\mathcal{M}^2$ TR. PROGRESSIVE [30]	2021	0.378	0.232	0.154	0.107	0.145	0.272	-
	Contrastive Attention [25]	2021	0.350	0.219	0.152	0.109	0.151	0.283	-
	AlignTransformer [53]	2021	0.378	0.235	0.156	0.112	0.158	0.283	-
	$\mathcal{M}^2$ Trans w/ NLL [28]	2021	-	-	-	0.105	-	-	0.445
	$\mathcal{M}^2$ Trans w/ NLL+BS+f_{C_E} [28]	2021	-	-	-	0.111	-	-	0.492
	$\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_{EN}}$ [28]	2021	-	-	-	0.114	-	-	**0.509**
	ITA [47]	2022	**0.395**	**0.253**	0.170	0.121	0.147	0.284	-
	CvT-212DistilGPT2 [29]	2022	0.392	0.245	0.169	0.124	0.153	**0.285**	0.361
	RGRG	Ours	0.373	0.249	**0.175**	**0.126**	**0.168**	0.264	0.495

Table 1. Natural language generation (NLG) metrics for the full report generation task. Our model is competitive with or outperforms previous state-of-the-art models on a variety of metrics. Dashed lines highlight the BLEU scores of the best baselines without processed (*i.e.* lowercased) reference reports (since lowercasing increases BLEU scores [35]). CIDEr score denoted by † cited from [28].

Dataset	Method	RL	Year	$P_{mic\text{-}5}$	$R_{mic\text{-}5}$	$F_{1,\,mic\text{-}5}$	$P_{ex\text{-}14}$	$R_{ex\text{-}14}$	$F_{1,\,ex\text{-}14}$
MIMIC-CXR	R2Gen [7]	✗	2020	0.412	0.298	0.346	0.331	0.224	0.228
	$\mathcal{M}^2$ Trans w/ NLL [28]	✗	2021	0.489	0.411	0.447	-	-	-
	$\mathcal{M}^2$ Trans w/ NLL+BS+f_{C_E} [28]	✓	2021	0.463	**0.732**	**0.567**	-	-	-
	$\mathcal{M}^2$ Trans w/ NLL+BS+$f_{C_{EN}}$ [28]	✓	2021	**0.503**	0.651	**0.567**	-	-	-
	CMN [6]	✗	2021	-	-	-	0.334	0.275	0.278
	Contrastive Attention [25]	✗	2021	-	-	-	0.352	0.298	0.303
	$\mathcal{M}^2$ TR. PROGRESSIVE [30]	✗	2021	-	-	-	0.240	0.428	0.308
	CvT-212DistilGPT2 [29]	✗	2022	-	-	-	0.359	0.412	0.384
	RGRG	✗	Ours	0.491	0.617	0.547	**0.461**	**0.475**	**0.447**

Table 2. Clinical efficacy (CE) metrics micro-averaged over 5 observations (denoted by mic-5) and example-based averaged over 14 observations (denoted by ex-14). RL represents reinforcement learning. Our model outperforms all non-RL models by large margins and is competitive with the two RL-based models directly optimized on CE metrics. Dashed lines highlight the scores of the best non-RL baseline. Micro-averaged results of R2Gen cited from [28], all example-based results are cited from [29].

Selection-based sentence generation. Quantitatively evaluating this task poses a challenge, since there is no strict ground-truth for sentences generated from manually drawn boxes. As such, we decided to evaluate this task by investigating the impact that randomly varying bounding boxes in relation to ground-truth bounding boxes of anatomical regions have on the METEOR scores (thereby simulating radiologists manually drawing bounding boxes around anatomical regions). Specifically, we vary bounding boxes by either position, aspect ratio, or scale, as illustrated in Fig. 4. Intuitively, we want to verify that small deviations from the ground-truth would have a negligible impact on the quality of the generated sentences, while large deviations would have a noticeable negative impact, thereby verifying that selection-based sentence generation is robust to different forms of drawn boxes while still being sensitive to the rough position of the selected region. We conduct several runs sampling from a normal distribution with increasingly higher standard deviations to vary the bounding boxes, thus achieving ever larger deviations from the ground-truth. We refer to the supplementary materials for details.

5. Results and discussion

5.1. Radiology report generation

Our model shows excellent radiology report generation performance (see Tab. 1 and Tab. 2), either being competitive with or outperforming previous models on conventional NLG metrics – setting a new state-of-the-art (SOTA) on the METEOR metric – as well as clinically relevant CE metrics – outperforming all models not directly optimized for these metrics by large margins.

On the BLEU scores, our model is competitive with the latest SOTA models [29, 47]. However, it is worth noting that [29] and [47] applied further processing such as lowercasing to the reference reports, which is well known to significantly increase BLEU scores [35] (see more details in supplementary materials). In comparison to the best baseline [28] without processed reference reports (highlighted by dashed lines), our model improves the BLEU-4 score by Δ+10.5%. On the METEOR score, which is case-insensitive, our model outperforms the latest SOTA model [29] by Δ+9.8% and achieves a Δ+6.3% increase

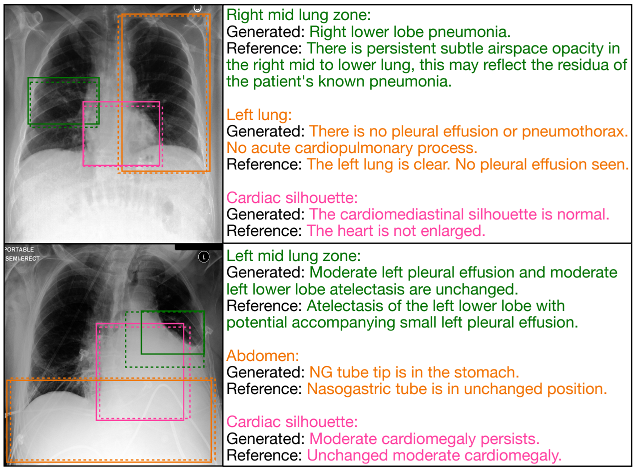

Figure 3. Qualitative results of anatomy-based sentence generation, including predicted (dashed boxes) and ground-truth (solid boxes) anatomical regions. Sentences are color-coded according to their corresponding anatomical regions. We observe that the model generates pertinent, anatomy-related sentences.

Region	RL	LL	SP	MED	CS	AB
METEOR	0.104	0.105	0.165	0.119	0.110	0.119

(a) Six prominent regions: *right lung* (RL), *left lung* (LL), *spine* (SP), *mediastinum* (MED), *cardiac silhouette* (CS) and *abdomen* (AB).

Region Subset	All	Normal	Abnormal	AS-Ratio
METEOR	0.115	0.202	0.064	1.938

(b) Micro-average over all, normal, and abnormal regions, respectively, and AS-Ratio.

Table 3. Language model results on the anatomy-based sentence generation task. We report (a) the per-anatomy METEOR scores for six prominent regions as well as (b) region subsets and the Anatomy-Sensitivity-Ratio (AS-Ratio). These results verify that our model generates anatomy-specific sentences as the AS-Ratio is two times higher than it would be for anatomy-agnostic sentences.

against the best baseline [53]. The score in the ROUGE-L (F1) metric, a summarization evaluation metric, is noticeably lower compared to the baselines. We believe that this can be attributed to the low precision score of the region selection module (see supplementary materials), *i.e.* more regions tend to be selected for a generated report than are described in the reference report (in turn causing a decrease in ROUGE-L precision).

For the micro-averaged CE metrics, our model achieves a substantial $+10.0\%$ increase ($\Delta{+}22.4\%$) in F1 score against the best baseline [28] (highlighted with dashed lines) not optimized via reinforcement learning (RL). Compared to the best baseline for the example-based CE metrics, our model again shows strong improvements with a $+6.3\%$ increase ($\Delta{+}16.4\%$) in F1 score. Our model is outperformed yet competitive with the baseline [28] explicitly optimized on CE metrics via RL, which demonstrates the effectiveness of our divide-and-conquer approach towards generating complete and consistent reports.

5.2. Anatomy-based sentence generation

Our model generates pertinent, anatomy-related sentences, as verified by quantitative (Tab. 3) and qualitative (Fig. 3) results.

We showcase generated sentences for various selected anatomies of two test set images in Fig. 3. Predicted bounding boxes closely align with the ground-truth in most cases, which is consistent with the good quantitative results of the object detector (we refer to the supplementary materials).

The anatomy-sensitivity observed in the qualitative results is further verified by the Anatomy-Sensitivity-Ratio of 1.938 in Tab. 3, which means that the METEOR score for sentences generated for a given anatomy is almost twice the score for when the model would have generated anatomy-agnostic sentences. It is noteworthy that generated descriptions of abnormal regions tend to reference earlier examinations (hence "*unchanged*" or "*persists*"), which can be explained by the sequential nature of examinations, especially if the progression of a disease is being tracked. Also, we observed that abnormal regions tend to have more diverse reference descriptions than normal regions. This in turn may cause lower scores in conventional NLG metrics for abnormal regions compared to normal regions (see Tab. 3), even though the generated description may be clinically accurate (see "*right mid lung zone*" in the upper image of Fig. 3).

5.3. Selection-based sentence generation

Our model shows high robustness towards deviations from the ground-truth in aspect ratio and scale while being sensitive to the position (see Fig. 5). This implies that when a region is specifically targeted for examination by manual bounding box annotation, the radiologist only has to position the bounding box close to the region of interest without worrying too much about capturing the exact aspect ratio and scale of the region. Thus, the integration of selection-based sentence generation as an interactive component into a clinical setting is viable.

Fig. 5 showcases this effect by plotting the change in METEOR score when manually drawn bounding boxes increasingly deviate from the ground-truth for different variation types (see Fig. 4). Since aspect ratio and scale are varied multiplicatively (with 1.0 representing no variation), the scores are plotted against ranges of multiplicative factors, whereas position is varied additively (with 0.0 representing no variation), thus the scores are plotted against increments of additive factors. We observe that the performance declines very slowly when the aspect ratio is varied and somewhat more when the scale is varied, while for position varia-

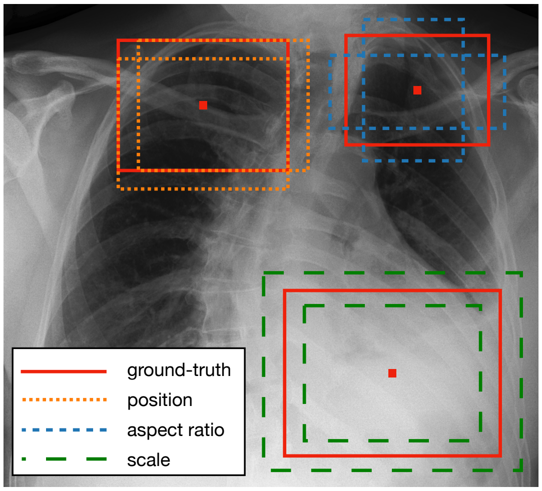

Figure 4. Illustration of the different bounding box variations used to evaluate the model's performance in generating selection-based sentences (*i.e.*, generated for manually drawn bounding boxes). We randomly vary the position, aspect ratio, and scale in order to simulate deviations of manually drawn bounding boxes from ground-truth boxes.

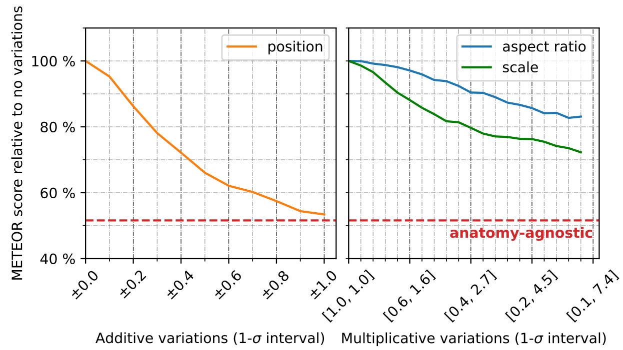

Figure 5. METEOR score of sentences generated from bounding boxes that increasingly deviate from the ground-truth to simulate manually drawn boxes (see Fig. 4). For reference, we also show the anatomy-agnostic METEOR score (red) for highlighting the threshold when anatomy-specificity is no longer achieved. Selection-based sentence generation is more sensitive to position and robust towards aspect ratio and scale, demonstrating the viability of this interactive application.

tions the model quickly reaches the score threshold at which generated sentences are considered anatomy-agnostic (see AS-Ratio presented in Tab. 3).

5.4. Component analysis and ablation study

In the supplementary materials, we provide detailed quantitative results and discussion on the object detector, the region selection, and the abnormality classification module. In addition, we conduct an ablation study on the latter two modules, verifying their effectiveness and their relevance for the overall model performance, and present further qualitative results.

6. Conclusion

In this paper, we present a simple yet effective approach to radiology report generation by explicitly focusing on salient anatomical regions through object detection and generating region-specific descriptions. Our method provides a high degree of explainability by visually grounding generated sentences on anatomical regions. Novel interactive capabilities allow radiologists direct involvement in the decision-making process (*e.g.*, by selecting a subset of anatomies or drawing bounding boxes for sentence generation). Such interactivity is crucial for successful integration into a real clinical setting. The experiments verify the effectiveness of our method for generating clinically accurate reports and offering these advantageous interactive capabilities. We hope that our work encourages further research towards region-guidance in radiology report generation.

Ethical Considerations. While automatic radiology report generation has the potential to substantially improve efficiency in the clinical workflow, erroneous diagnostic evaluations by such systems can have direct, harmful consequences for patients. As the clinical accuracy of automatic diagnostic methods improves, we see an increased risk of overreliance [33] by automation bias (tendency to favor machine-generated decisions while disregarding contrary data [11]). Interestingly, detailed explanations provided by a system may even exacerbate automation bias [42]. However, we believe that our approach of explainability through visual grounding and additional interactive capabilities encourages manual verification and intervention by radiologists, which may mitigate the risk of overreliance.

Limitations. While technically simple, our method requires strong supervision, currently only provided by the Chest ImaGenome dataset, and is thus hard to generalize to other report generation tasks. Adapting our approach for limited supervision is a promising future research direction. Secondly, our method only considers chest X-rays in isolation, whereas chest X-rays are usually diagnostically evaluated in the context of previous examinations. The Chest ImaGenome dataset contains localized comparison relations for anatomical regions across sequential exams, and incorporating this information has the potential to improve the clinical accuracy of generated reports. Finally, the reference reports of the MIMIC-CXR dataset contain sentences describing objects that cannot be attributed to an anatomical region (*e.g.* surgical clips) and are thus not covered by our method. Future work may thus consider a hybrid system utilizing region- and image-level visual features to generate both region- and such image-level sentences.

References

[1] Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *IMU*, 24:100557, 2021. 2, 4

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[3] Sarah Bastawrous and Benjamin Carney. Improving patient safety: avoiding unread imaging exams in the national va enterprise electronic health record. *Journal of digital imaging*, 30(3):309–313, 2017. 1

[4] William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. Baselines for chest x-ray report generation. In *ML4H Workshop*, pages 126–140, 2020. 5

[5] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*, pages 9962–9971, 2020. 3

[6] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *ACL*, pages 5904–5914, 2021. 2, 5, 6

[7] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, pages 1439–1449, 2020. 2, 5, 6

[8] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, pages 8307–8316, 2019. 3

[9] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020. 2, 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 3

[11] J. Raymond Geis, Adrian P. Brady, Carol C. Wu, Jack Spencer, Erik Ranschaert, Jacob L. Jaremko, Steve G. Langer, Andrea Borondy Kitts, Judy Birch, William F. Shields, et al. Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. *Radiology*, 293(2):436–440, 2019. 1, 2, 8

[12] Stacy K. Goergen, Felicity J. Pool, Tari J. Turner, Jane E. Grimm, Mark N. Appleyard, Carmel Crock, Michael C. Fahey, Michael F. Fay, Nicholas J. Ferris, Susan M. Liew, et al. Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *JMIRO*, 57(1):1–7, 2013. 1

[13] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. 5

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[16] Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *ACL*, pages 6570–6580, 2019. 2

[17] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *ACL*, pages 2577–2586, 2018. 2

[18] Alistair E. W. 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):1–8, 2019. 2, 5

[19] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr database (version 2.0.0). *PhysioNet*, 2019. 1, 2, 5

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, pages 4565–4574, 2016. 2

[21] Xiangyang Li, Shuqiang Jiang, and Jungong Han. Learning object context for dense captioning. In *AAAI*, volume 33, pages 8650–8657, 2019. 2

[22] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *NIPS*, 31, 2018. 2

[23] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[24] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, pages 13753–13762, 2021. 2, 5, 6

[25] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. In *ACL-IJCNLP*, pages 269–280, 2021. 2, 5, 6

[26] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *MLHC*, pages 249–269, 2019. 2, 5

[27] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. 1, 2

[28] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *NAACL*, pages 5288–5304, 2021. 1, 5, 6, 7

[29] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm-starting. *arXiv preprint arXiv:2201.09405*, 2022. 2, 5, 6

[30] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In *EMNLP*, pages 2824–2832, 2021. 2, 5, 6

[31] Yannis Papanikolaou and Andrea Pierleoni. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*, 2020. 4

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 5

[33] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai: Literature review. *Technical report, Microsoft Research*, 2022. 8

[34] Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. Inspecting state of the art performance and nlp metrics in image-based medical report generation. *arXiv preprint arXiv:2011.09257*, 2020. 5

[35] Matt Post. A call for clarity in reporting bleu scores. In *WMT*, pages 186–191, 2018. 6

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[37] Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C. Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012. 1

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 28, 2015. 3

[39] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020. 2

[40] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ*, 359, 2017. 1

[41] Andrew B. Rosenkrantz, Danny R. Hughes, and Richard Duszak Jr. The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279(1):175–184, 2016. 1

[42] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: expertise and explanations. In *ACM IUI*, pages 240–251, 2019. 8

[43] Zhuang Shao, Jungong Han, Demetris Marnerides, and Kurt Debattista. Region-object relation-aware dense captioning via transformer. *IEEE Trans Neural Netw Learn Syst.*, 2022. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017. 2

[45] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 5

[46] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 2

[47] Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. An inclusive task-aware framework for radiology report generation. In *MICCAI*, pages 568–577, 2022. 2, 5, 6

[48] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, pages 9049–9058, 2018. 2

[49] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, et al. Chest imagenome dataset for clinical reasoning. In *NIPS*, 2021. 5

[50] 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 (version 1.0.0). *PhysioNet*, 2021. 5

[51] 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 *ICML*, pages 2048–2057, 2015. 2

[52] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. Context and attribute grounded dense captioning. In *CVPR*, pages 6241–6250, 2019. 2

[53] 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 *MICCAI*, pages 72–82. Springer, 2021. 2, 5, 6, 7

[54] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016. 2

[55] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019. 4

[56] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *AAAI*, volume 34, pages 12910–12917, 2020. 2

[57] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. In *CVPR*, pages 8395–8404, 2019. 3

[58] Zachary M. Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M. Rush. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*, 2019. 4