Textual Inversion and Self-supervised Refinement for Radiology Report Generation

Yuanjiang Luo[1*], Hongxiang Li[2*], Xuan Wu[3], Meng Cao[4], Xiaoshuang Huang[5], Zhihong Zhu[2], Peixi Liao[6], Hu Chen[3†], and Yi Zhang[7]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Sichuan, China
[2] Peking University, Shenzhen, China
[3] College of Computer Science, Sichuan University, Sichuan, China
luoyj@stu.scu.edu.cn, huchen@scu.edu.cn
[4] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[5] China Agricultural University, Beijing, China
[6] The Sixth People's Hospital of Chengdu, Sichuan, China
[7] School of Cyber Science and Engineering, Sichuan University, Sichuan, China

Abstract. Existing mainstream approaches follow the encoder-decoder paradigm for generating radiology reports. They focus on improving the network structure of encoders and decoders, which leads to two shortcomings: overlooking the modality gap and ignoring report content constraints. In this paper, we proposed **T**extual **I**nversion and **S**elf-supervised **R**efinement **(TISR)** to address the above two issues. Specifically, textual inversion can project text and image into the same space by representing images as pseudo words to eliminate the cross-modeling gap. Subsequently, self-supervised refinement refines these pseudo words through contrastive loss computation between images and texts, enhancing the fidelity of generated reports to images. Notably, **TISR** is orthogonal to most existing methods, plug-and-play. We conduct experiments on two widely-used public datasets and achieve significant improvements on various baselines, which demonstrates the effectiveness and generalization of **TISR**. The code will be available soon.

Keywords: Radiology report generation · Cross-modal learning · Textual inversion · Auxiliary diagnosis

1 Introduction

Radiology report generation provides the basis for physician diagnosis [9]. However, observing radiograph and writing report is time-consuming and laborious for doctors [21]. It's even error-prone for inexperienced doctors as they often struggle to accurately capture the abnormalities in images [2, 20]. Previous approaches adopt the framework of image captioning [8] straightforwardly and

* Equal contributions.
† Corresponding author.

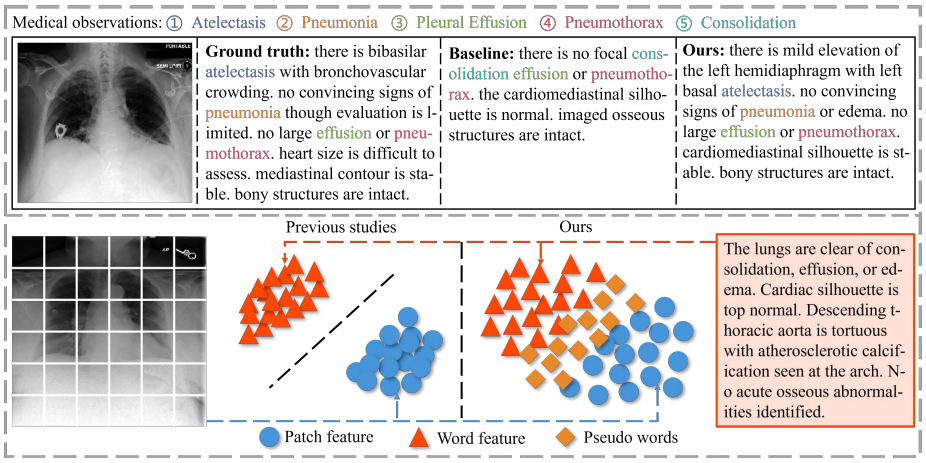

Fig. 1. Existing challenges in radiology report generation.

make it more suitable for generating radiology reports by improving image encoders [27, 16, 29] to be better adapted to medical images or refining text decoders to generate long paragraphs [30, 14, 10]. Building upon this, innovative techniques are utilized to improve performance, such as knowledge graph [32], causal inference [3], and dynamic graph [15].

Despite these notable advances, there are still two challenges in generating accurate reports. **(1)** Existing methods cannot explicitly constrain the reports generated by the text decoder to be faithful to visual information. Prior method [4] suffers from the phenomenon of illusion generation, as shown in Fig. 1. It generates "*consolidation*", which is not mentioned in the ground truth while misses "*atelectasis*" and "*pneumonia*". Some works have enhanced the ability of grounding by extracting additional expert information, such as anchor box [27] and sentence retrieval library [13]. However, their implementation needs to make additional labels or reconstruct the entire dataset, which not only requires expensive costs but is not always accessible in clinical applications. **(2)** The inherent modal gap between medical images and reports. Previous approaches [27, 16, 29, 30, 14, 10, 31] adhere to the image encoder-text decoder paradigm [19], which lacks cross-modal interaction. As image and text exist in distinct feature spaces with a feature gap between them, we propose to fill this gap with pseudo words, constructing a unified public hidden space for image and text, as shown in Fig. 1.

We propose **T**extual **I**nversion and **S**elf-supervised **R**efinement **(TISR)** to solve the problems discussed above. We employ a lightweight mapping module, named textual inversion, to convert image features into text features [25]. Through textual inversion, the pseudo words obtained by transforming image embeddings contain both image features and linguistic spatial characteristics. Textual inversion can eliminate the spatial gap effectively, making the features of two modalities be computed in the common compact space. We then perform

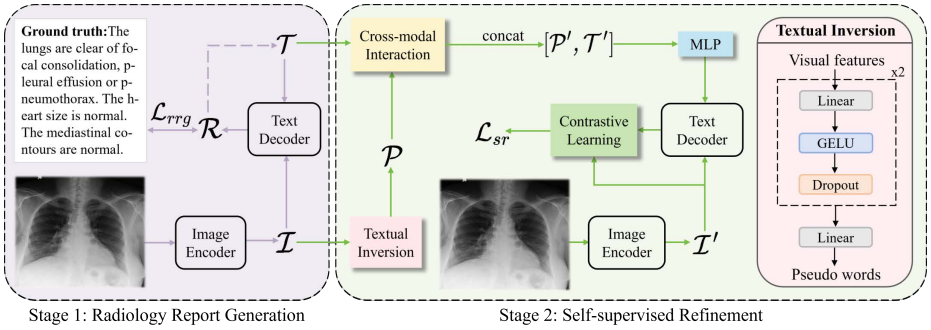

Fig. 2. Overview of our method. The arrow dashed line indicates that before obtaining the entire report, the current word is generated by the image features and text embeddings obtained by encoding the previously generated words.

self-supervised refinement by calculating contrastive loss between the obtained pseudo words and image features. Instead of relying on ground truth, **TISR** guides the network to generate reports faithful to the images by minimizing the contrastive loss. Experimental results on two widely used datasets and three radiology report generation networks verify the efficacy and plug-and-play capability of **TISR**. In summary, the contributions of this paper are as follows:

- We bridge the modality gap by transforming visual features into linguistic space through textual inversion.
- The self-supervised refinement module searches for text representations close to the image content to minimize the contrastive loss. Consequently, we can generate faithful reports to radiographs, providing more credible diagnostic information for clinical practice.
- Our **TISR** is orthogonal to other radiology report generation networks, plug-and-play. Experimental results show that by improving the network with **TISR**, the accuracy improved compared to the baselines.

2 Method

As shown in Fig. 2, our pipeline consists of two stages. We extract image features $\mathcal{I} \in \mathbb{R}^{B\times M\times D}$ from a radiograph through an image encoder [11], where B is the batch size, M is the number of the processed patches and D is the dimension. With the image features $\mathcal{I}$ and the previously generated text embeddings $\mathcal{T}_{n-1}$, the text decoder [6] can obtain the log probability $\mathcal{O}_n$ of the next word.

$$\mathcal{O}_n = f_d(\mathcal{I}, \mathcal{T}_{n-1}) \quad \text{and} \quad \mathcal{T}_{n-1} = f_e(\mathcal{R}_{n-1}), \tag{1}$$

where f_e, f_d and n denote text encoder, text decoder and one word of the target report respectively. By applying linear and softmax to the log probability, we

obtain the word in the vocabulary corresponding to the highest probability and use it as the n-th word of the report.

$$\mathcal{R}_n = \text{softmax}(f_l(\mathcal{O}_n)). \tag{2}$$

where f_l is a linear layer that maps the outputs of decoder to the vocabulary. We can finally obtain the complete text embeddings $\mathcal{T} \in \mathbb{R}^{B\times N\times D}$ and report $\mathcal{R} \in \mathbb{R}^{B\times N}$ after continuous autoregression of the text decoder [22], where N represents the length of the target sequence.

Image features $\mathcal{I}$ are processed by textual inversion to generate pseudo words $\mathcal{P}$. In the self-supervised refinement, we supervise the network to generate more refined pseudo words by calculating the contrastive loss between text features and image features instead of using ground truth as the supervision signal. The details are illustrated in the following subsections.

2.1 Textual Inversion

Radiology report generation is an image-to-text cross-modal task, as medical images and radiology reports are in two different feature spaces. Existing methods are more tend to improve the overall performance by extracting refined image features [27, 16, 29] or improving the network structure of the text decoder [30, 14, 10, 31] and ignoring the gap between modalities. Therefore, we propose textual inversion to reconstruct image representation within the text embedding space to eliminate the spatial gap. In this module, we map image embeddings $\mathcal{I}$ to pseudo words $\mathcal{P} \in \mathbb{R}^{B\times M\times D}$ via feeding image features into a three-layered full-connected network, which can be formulated as:

$$\mathcal{P} = f_l(\text{Dropout}(\text{GELU}(f_l(P_1)))) \quad \text{and} \quad P_1 = \text{Dropout}(\text{GELU}(f_l(\mathcal{I}))). \tag{3}$$

2.2 Self-supervised Refinement

After obtaining the pseudo words, we input them into the text decoder after a series of operations to obtain $\mathcal{O}'$. We explicitly constrain that the generated pseudo words should be able to represent the image features sufficiently by calculating the contrastive loss between $\mathcal{O}'$ and the image feature $\mathcal{I}'$. This optimization process guides the network to generate reports that are faithful to images.

We first perform cross-modal interaction by employing cross attention mechanism [28] between pseudo words $\mathcal{P}$ and text embeddings $\mathcal{T}$. Since pseudo words are derived directly from image features, it is beneficial to align visual and linguistic features through this interaction. This process can be expressed as:

$$\mathcal{P}' = \text{softmax}(\frac{\mathcal{P}\mathcal{T}^T}{\sqrt{D}})\mathcal{T}, \tag{4}$$

$$\mathcal{T}' = \text{softmax}(\frac{\mathcal{T}\mathcal{P}^T}{\sqrt{D}})\mathcal{P}, \tag{5}$$

We assume that the pseudo words can compensate for the missing information or correct the redundant information for the text feature $\mathcal{T}$. Based on this intuition, we concatenate aligned text feature $\mathcal{T}'$ with aligned pseudo words $\mathcal{P}'$. The concatenated features can be fused well through a multi-layer perceptron (MLP), which contains two linear layers and ReLU activation function. Thus, we obtained the processed pseudo words $\mathcal{P}'' \in \mathbb{R}^{B\times S\times D}$, where $S = M + N$. It can be expressed using a formula:

$$\mathcal{P}'' = \mathrm{MLP}[\mathcal{P}', \mathcal{T}']. \tag{6}$$

The log probability $\mathcal{O}' \in \mathbb{R}^{B\times S\times D}$ is obtained by decoding $\mathcal{P}''$ and $\mathcal{I}'$. We then implement self-supervised refinement by calculating contrastive loss between text embeddings $\mathcal{O}'$ and image features $\mathcal{I}'$. By minimizing the contrastive loss, we encourage the network to generate $\mathcal{P}''$ that closely resemble the expression of $\mathcal{I}'$. After continuous back-propagation and optimization, the generated pseudo words $\mathcal{P}$ can adequately represent the image semantics, which is beneficial for the generation of reports faithful to the original images.

2.3 Training Objective

We utilize $\mathcal{L}_{rrg}$ to quantify the difference between the generated report and the ground truth [5], thus guiding the model to generate reports that are close to the ground truth. The formulation of $\mathcal{L}_{rrg}$ is as follows:

$$\mathcal{L}_{rrg} = -\frac{1}{\sum M}\sum_{b=1}^{B}\sum_{s=1}^{S} M_{bs} O_{bs,T_{bs}}. \tag{7}$$

The log probability of the output against the target sequence T is obtained by $O_{bs,T_{bs}}$ for the position s of the b-th sample. To ensure consistent input sequence length, all sequences are filled to the same length. The mask M_{bs} indicates whether a real word exists at the position: if present, it is 1; otherwise, it is 0. The log probability of the filled part is set to 0 by multiplying with the mask M_{bs} to prevent the filled part from affecting $\mathcal{L}_{rrg}$. Finally, we normalize $\mathcal{L}_{rrg}$ by dividing it with the sum of all mask values $\sum M$ to ensure that the loss value is not affected by changes in sequence length.

In addition to optimizing the network to generate more accurate reports through $\mathcal{L}_{rrg}$, we also constrain the textual inversion to generate pseudo words that are close to the image representation through $\mathcal{L}_{sr}$. We obtained the score matrix $\mathcal{S}$ by calculating the correlation between image features and text features via dot product, which can be expressed as $\mathcal{S} = \mathcal{I}' \times {\mathcal{O}'}^{T}$, where $\times$ denotes matrix multiplication. we evaluate the cosine similarity between image features and text features and get the score matrix $\mathcal{S}$ of size $B \times B$. We optimize the network by constructing a symmetric cross-entropy loss to maximize the cosine similarity between B real image-text pairs while minimizing the cosine similarity between $B^2 - B$ unpaired image-text pairs [24].

$$\mathcal{L}_{sr} = -\frac{1}{2}\left(\sum_{b=1}^{B}\mathcal{M}_b log(S_b) + \sum_{b=1}^{B}\mathcal{M}_b log({S_b}^{T})\right). \tag{8}$$

$\mathcal{M}$ is a matrix of size $B \times B$, where the elements on the diagonal are 1, indicating positive samples, while the off-diagonal elements are 0, indicating negative samples. The overall loss function $\mathcal{L}$ of our network is defined as: $\mathcal{L} = \mathcal{L}_{rrg} + \mathcal{L}_{sr}$. Instead of relying on manually labeled datasets, we leverage contrastive learning to measure the similarity between text and image, guiding the network to optimize parameters for generating reports faithful to the visual content.

3 Experiment

3.1 Dataset and Evaluation Metrics

Dataset. We conducted experiments on two widely-used datasets: the small dataset IU X-ray† [7] (containing 7,470 chest X-ray images and 3,955 corresponding reports) and the large dataset MIMIC-CXR† [12] (containing 377,110 images and 227,835 corresponding reports). To ensure consistency and fairness in comparisons, we followed the data processing methods utilized by the three baselines [3–5]. After excluding samples without corresponding radiology reports, IU X-ray is divided into training, validation and testing sets with a proportion of 7:1:2 [17] while MIMIC-CXR is divided according to the official splits [5].

Evaluation Metrics. We evaluate **TISR** on natural language generation (NLG) metrics including BLEU [23], METEOR [1] and ROUGE-L [18], which are widely used to assess the fluency and accuracy of generated reports. We not only focus on the quality of the generated reports but also on their ability to accurately capture lesions in the images. Therefore, we employ clinical efficacy (CE) metrics to evaluate the detection accuracy of generated reports. CheXbert [26]is widely used as an automated deep-learning based chest radiology report labeler that can label for 14 medical observations [3–5, 13, 15, 17, 22]. We extract labels from both the generated reports and the ground truth using CheXbert, and then compare these labels to obtain precision, recall, and F1.

3.2 Experiments Results and Analyses

Comparison with Baselines. To verify the generalization and effectiveness of **TISR**, we use R2Gen [5], R2GenCMN [4] and VLCI [3] as the baseline models in our experiments. These baseline models are improved with **TISR**, and the results are compared with the original baselines, as shown in Table 1. Experimental results show that all metrics are enhanced by improving networks with **TISR**, which indicates that **TISR** can eliminate the gap between modalities and generate more accurate reports. It is remarkable that our approach does not require additional data and can seamlessly integrate into these baselines [3–5], which is of great importance for network migration and practical applications. What's more, we can recognize from the results that prior methods have overlooked the impact of the gap between modalities on radiology report generation. Hence, future research should focus more on cross-modal interactions.

† https://openi.nlm.nih.gov/.

† https://physionet.org/content/mimic-cxr/2.0.0/.

Table 1. Comparison between baselines and the improved network with **TISR**. △ denotes the improvements compared to the baselines. * denotes our re-implementation of baselines. MTR and RG-L denote METEOR and ROUGE-L, respectively.

Method	NLG Metrics						CE Metrics		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	MTR	RG-L	Precision	Recall	F1
Experimental results on IU X-ray dataset.									
R2Gen* [5]	0.443	0.286	0.212	0.168	0.175	0.355	-	-	-
+TISR(Ours)	0.470	0.310	0.233	0.187	0.194	0.369	-	-	-
△	**+0.027**	**+0.024**	**+0.021**	**+0.019**	**+0.019**	**+0.014**	-	-	-
R2GenCMN* [4]	0.469	0.300	0.215	0.164	0.190	0.370	-	-	-
+TISR(Ours)	0.483	0.313	0.229	0.176	0.191	0.371	-	-	-
△	**+0.014**	**+0.013**	**+0.014**	**+0.012**	**+0.001**	**+0.001**	-	-	-
VLCI* [3]	0.467	0.306	0.225	0.175	0.193	0.377	-	-	-
+TISR(Ours)	0.485	0.318	0.232	0.179	0.199	0.382	-	-	-
△	**+0.018**	**+0.012**	**+0.007**	**+0.004**	**+0.006**	**+0.005**	-	-	-
Experimental results on MIMIC-CXR dataset.									
R2Gen* [5]	0.350	0.214	0.143	0.103	0.135	0.271	0.424	0.254	0.317
+TISR(Ours)	0.358	0.219	0.147	0.106	0.139	0.275	0.467	0.302	0.367
△	**+0.008**	**+0.005**	**+0.004**	**+0.003**	**+0.004**	**+0.004**	**+0.043**	**+0.048**	**+0.050**
R2GenCMN* [4]	0.344	0.210	0.139	0.098	0.136	0.275	0.455	0.317	0.374
+TISR(Ours)	0.363	0.224	0.149	0.105	0.143	0.279	0.450	0.344	0.390
△	**+0.019**	**+0.014**	**+0.010**	**+0.007**	**+0.007**	**+0.004**	-0.005	**+0.027**	**+0.016**
VLCI* [3]	0.393	0.239	0.159	0.113	0.150	0.276	0.439	0.283	0.344
+TISR(Ours)	0.396	0.242	0.161	0.115	0.149	0.278	0.453	0.306	0.366
△	**+0.003**	**+0.003**	**+0.002**	**+0.002**	**+0.001**	**+0.002**	**+0.014**	**+0.023**	**+0.022**

Table 2. Ablation experiment of **TISR**.

Textual Inversion	Self-supervised Refinement	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
		0.443	0.286	0.212	0.168	0.175	0.355
	✓	0.462	0.304	0.224	0.176	0.190	0.361
✓		0.463	0.289	0.206	0.157	0.181	0.356
✓	✓	**0.470**	**0.310**	**0.233**	**0.187**	**0.194**	**0.369**

Ablation Study. To explore the effectiveness of each component in **TISR** and the rationality of the network structure, we conducted various ablation experiments. First of all, we explored the effectiveness of textual inversion and self-supervised refinement, as shown in Table 2. The significance of textual inversion is investigated by calculating the contrastive loss between image features $\mathcal{I}$ and text embeddings $\mathcal{T}$, while the role of self-supervised refinement is explored through the calculation of the contrastive loss between image features $\mathcal{I}'$ and pseudo words $\mathcal{P}$.

Secondly, we carried out experiments to explore the structure of **TISR**. We replace the three-layer linear structure with a three-layer transformer encoder to explore the structure of the textual inversion module. It's easy to see that the result is worse than MLP with the same dimension of the hidden layer from Table 3. We hypothesize that this is because the medical image features are not complex,

Table 3. Ablation experiments on the structure of textual inversion.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Transformer	0.416	0.268	0.194	0.149	0.178	0.350
MLP	**0.470**	**0.310**	**0.233**	**0.187**	**0.194**	**0.369**

Table 4. Ablation experiments on the structure of self-supervised refinement.

Text Decoder	Pseudo Words	Cross-modal Interaction	MLP	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
	✓	✓	✓	0.442	0.277	0.198	0.152	0.176	0.352
✓		✓	✓	0.446	0.291	0.218	0.175	0.183	0.365
✓	✓		✓	0.459	0.298	0.221	0.173	0.186	0.358
✓	✓	✓		0.432	0.301	0.226	0.175	0.190	**0.398**
✓	✓	✓	✓	**0.470**	**0.310**	**0.233**	**0.187**	**0.194**	0.369

so using a transformer may lead to overfitting, and it will also increase computational costs. We investigate the significance of cross-modal interaction and MLP by removing cross attention and MLP from the complete self-supervised refinement network respectively. Pseudo words' significance is investigated by directly incorporating $\mathcal{T}$ into the self-supervised refinement network. Furthermore, the importance of decoding text embeddings is explored by computing the contrastive loss between $\mathcal{P}''$ and $\mathcal{I}'$. We can speculate that each module plays an important role in generating more refined pseudo words from Table 4 since removing any one of them degrades the performance of the network.

Quantitative Analysis. We draw attention maps to explore the region of the medical image that the word of the generated report is interested in. Fig. 3 illustrates that the model improved by **TISR** is more sensitive to the correct regions and can generate reports that are closer to the ground truth. This demonstrates that our model can eliminate the cross-modal gap and thus generate reports faithful to images.

4 Conclusion

In this study, we propose **T**extual **I**nversion and **S**elf-supervised **R**efinement **(TISR)** to address the radiology report generation problem. By inverting image features into pseudo words, textual inversion aims to bridge the modality gap by representing visual features in the linguistic space. We employ self-supervised refinement to iteratively improve the quality of pseudo words by minimizing the contrastive loss between them and the image features. This iterative process helps to generate radiology reports that are faithful to the radiology image. **TISR** is designed to compensate for most existing approaches seamlessly, offering a plug-and-play solution. Significant improvements across all three baselines illustrate the effectiveness and generation of our proposed method.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

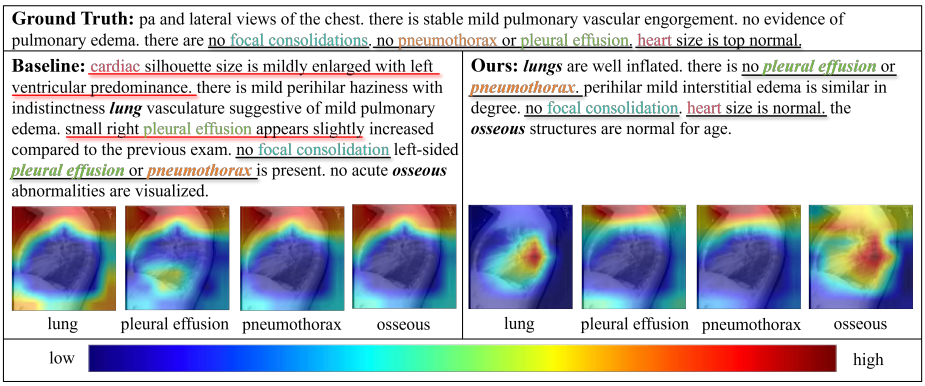

Fig. 3. Visualization. Red: the network is highly concerned about this area, blue: the area that is not concerned, black line: correct description, red line: incorrect description.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Brady, A., Laoide, R.Ó., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. The Ulster medical journal **81**(1), 3 (2012)
3. Chen, W., Liu, Y., Wang, C., Li, G., Zhu, J., Lin, L.: Visual-linguistic causal intervention for radiology report generation. arXiv preprint arXiv:2303.09117 (2023)
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
8. Deng, C., Ding, N., Tan, M., Wu, Q.: Length-controllable image captioning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 712–729. Springer (2020)
9. Elliott, J.: The value of case reports in diagnostic radiography. Radiography **29**(2), 416–420 (2023)
10. Harzig, P., Chen, Y.Y., Chen, F., Lienhart, R.: Addressing data bias problems for chest x-ray image report generation. arXiv preprint arXiv:1908.02123 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

12. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
13. Kong, M., Huang, Z., Kuang, K., Zhu, Q., Wu, F.: Transq: Transformer-based semantic query for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 610–620. Springer (2022)
14. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6666–6673 (2019)
15. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
16. Li, M., Liu, R., Wang, F., Chang, X., Liang, X.: Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web **26**(1), 253–270 (2023)
17. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018)
18. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Liu, C., Tian, Y., Song, Y.: A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199 (2023)
20. Manning, D., Ethell, S., Donovan, T., Crawford, T.: How do radiologists do it? the influence of experience and training on searching for chest nodules. Radiography **12**(2), 134–142 (2006)
21. McGaghie, W.C.: Education for chest radiograph interpretation performance improvement. Chest **164**(2), e57 (2023)
22. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19305–19314 (2023)
26. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv:2004.09167 (2020)
27. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023)

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Wang, Y., Wang, K., Liu, X., Gao, T., Zhang, J., Wang, G.: Self adaptive global-local feature enhancement for radiology report generation. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2275–2279. IEEE (2023)
30. Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 457–466. Springer (2018)
31. Yan, B., Pei, M., Zhao, M., Shan, C., Tian, Z.: Prior guided transformer for accurate radiology reports generation. IEEE Journal of Biomedical and Health Informatics **26**(11), 5631–5640 (2022)
32. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)