Structural Entities Extraction and Patient Indications Incorporation for Chest X-ray Report Generation

Kang Liu[1,2,3], Zhuoqi Ma[1,2,3,4] (✉), Xiaolu Kang[1,2,3], Zhusi Zhong[5], Zhicheng Jiao[4], Grayson Baird[4], Harrison Bai[6], and Qiguang Miao[1,2,3] (✉)

[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] Xi'an Key Laboratory of Big Data and Intelligent Vision, Xi'an, China
[3] Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xi'an, China
[4] Warren Alpert Medical School, Brown University, Providence, USA
[5] School of Electronic Engineering, Xidian University, Xi'an, China
[6] Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, Baltimore, USA
zhuoqi_ma@hotmail.com; qgmiao@xidian.edu.cn

Abstract. The automated generation of imaging reports proves invaluable in alleviating the workload of radiologists. A clinically applicable reports generation algorithm should demonstrate its effectiveness in producing reports that accurately describe radiology findings and attend to patient-specific indications. In this paper, we introduce a novel method, **S**tructural **E**ntities extraction and patient indications **I**ncorporation (SEI) for chest X-ray report generation. Specifically, we employ a structural entities extraction (SEE) approach to eliminate presentation-style vocabulary in reports and improve the quality of factual entity sequences. This reduces the noise in the following cross-modal alignment module by aligning X-ray images with factual entity sequences in reports, thereby enhancing the precision of cross-modal alignment and further aiding the model in gradient-free retrieval of similar historical cases. Subsequently, we propose a cross-modal fusion network to integrate information from X-ray images, similar historical cases, and patient-specific indications. This process allows the text decoder to attend to discriminative features of X-ray images, assimilate historical diagnostic information from similar cases, and understand the examination intention of patients. This, in turn, assists in triggering the text decoder to produce high-quality reports. Experiments conducted on MIMIC-CXR validate the superiority of SEI over state-of-the-art approaches on both natural language generation and clinical efficacy metrics. The code is available at https://github.com/mk-runner/SEI.

Keywords: Chest X-ray report generation · Structural entities extraction · Patient-specific indications · Cross-modal fusion · Similar historical cases.

1 Introduction

Radiology reports play a crucial role in delivering clear, accurate, and easily understandable medical information, thereby facilitating effective communication between doctors and patients. Nevertheless, this task is both highly specialized and time-consuming. Additionally, variations in proficiency, experience, and individual habits among radiologists would inevitably impact the quality and consistency of reports. Fortunately, the rapid evolution of artificial intelligence techniques, particularly deep learning [8,9,23], has significantly propelled the advancement of chest X-ray report generation (CXRG).

Currently, there is a considerable body of studies on medical report generation [5,32,10,25,26], contributing significantly to the improvement of clinical effectiveness and linguistic fluency in reports. However, two challenges persist in CXRG: 1) To truly meet clinical needs, report generation processes should incorporate patient-specific indications, such as previous treatment history or responses to specific diagnostic requirements, which cannot be derived exclusively from medical images. 2) Existing methods face challenges in effectively focusing on the cross-modal alignment between medical images and reports. This is attributed to the practice of assigning equal weights to presentation-style elements (e.g., sentence structure and grammar) and factual vocabulary (e.g., findings) in reports. Unfortunately, this limitation impacts their clinical efficacy.

In response to the challenges above, previous studies have made specific attempts. Concerning challenge 1): [11] and [21] leverage BiLSTM to encode the indication section with specific terms, facilitating the generation of purposeful reports. [17] adopts the LLaMA model [22] to generate reports based on indications and predicted positive conditions, falling short in fully exploiting valuable information within medical images. Furthermore, many existing works [4,15,30] treat the CXRG task as an image-to-text generation problem, neglecting the effect of patient-specific indications on CXRG. Regarding challenge 2): Numerous studies directly utilize reports and medical images for cross-modal alignment at various granularities (e.g., instance-level [28,31], sentence-level [6], and token-level [24]). However, these methods treat presentation-style elements and factual vocabulary equally in reports, potentially affecting the quality of cross-modal alignment and, consequently, the clinical efficacy of the generated reports. In light of this limitation, building directly upon the outcomes of RadGraph [12], KAD [31] achieves instance-wise cross-modal alignment between images and the factual vocabulary in reports, while [27] focuses on learning the mapping relationship between them using an encoder-decoder framework. Nevertheless, both methods overlook the noise and redundancy in the RadGraph outcomes.

In this paper, we introduce a novel method, **S**tructural **E**ntities extraction and patient indications **I**ncorporation (SEI), for chest X-ray report generation. SEI involves two stages: training the cross-modal alignment module and training the report generation module. In the first stage, SEI employs the structural entities extraction (SEE) approach to eliminate presentation-style vocabulary in reports and enhance the quality of factual entity sequences. Subsequently, a cross-modal alignment module is introduced between X-ray images and factual

entity sequences in reports, ensuring that the extracted image features implicitly preserve semantic similarity with their corresponding reference reports. In the second stage, leveraging the pre-trained model from the first stage, we conduct a gradient-free retrieval of similar historical cases for each sample from the training set. Following this, the cross-modal fusion network is deployed to integrate these cases, patient-specific indications, and imaging information. This process enables the text decoder to assimilate historical diagnostic information from similar cases, understand the examination intention of patients (e.g., symptoms), and attend to discriminative features of X-ray images. Finally, this contributes to triggering the text decoder for the generation of high-quality reports. The effectiveness of SEI is successfully validated on MIMIC-CXR in both specific and general scenarios, outperforming multiple state-of-the-art methods.

In summary, our key contributions are as follows: 1) We develop a structural entities extraction approach to extract factual entity sequences from reports. This step reduces the noise in the cross-modal alignment process, facilitating gradient-free retrieval of similar historical cases from the training set. 2) We introduce a cross-modal fusion network to integrate similar historical cases, patient-specific indications, and imaging information. This allows the text decoder to attend to discriminative features of X-ray images, assimilate historical empirical information from similar cases, and understand the examination intention of patients. 3) Experiments on MIMIC-CXR demonstrate that our SEI achieves state-of-the-art performance across almost all metrics. This highlights the capability of our model to generate reports with encouraging clinical efficacy and linguistic fluency.

2 Method

Our objective is to train a model capable of generating a report for a given X-ray image, conditioned on similar historical cases and patient-specific indications. As shown in Fig. 1, the model comprises two stages: pre-training, which involves a cross-modal alignment module enhanced by factual entity sequences, and fine-tuning, which is a report generation module based on both similar historical cases and patient-specific indications.

2.1 Cross-modal Alignment Module Enhanced by Factual Entity Sequences

Structural entities extraction approach for extracting factual entity sequences. Motivated by [27], we devise the structural entities extraction (SEE) approach. More precisely, we first extract entities from reports using RadGraph [12]. Afterward, we remove entities spanning two sentences, for example, "*in place . Swan Ganz*". For redundant entities at the same location, such as "*1.9 × 1.0 cm*" and "*1.0 cm*", we retain only the longest entity, specifically, the former. Following this, all remaining entities in the report are organized in their original order and divided into multiple factual entity subsequences, such as "*AICD in*

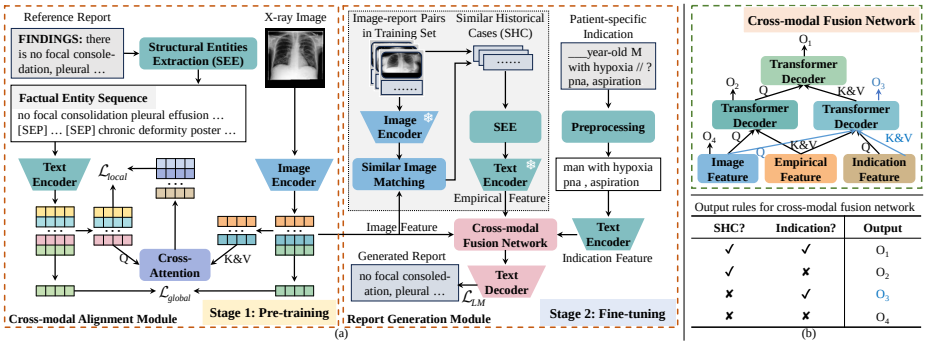

Fig. 1. Illustration of our SEI and cross-modal fusion network. (a) Overview of SEI, featuring dual encoders for extracting uni-modal features and a text decoder for report generation using X-ray images, similar historical cases (SHC), and patient-specific indications. The training paradigm of SEI includes 1) pre-training via the cross-modal alignment module; 2) gradient-free retrieval of similar historical cases using the pre-trained model from step 1), shown in the light grey box; 3) fine-tuning using the report generation module. (b) Details and output rules of the cross-modal fusion network.

place", using sentence periods. Notably, when a factual entity subsequence contains an "*OBS-DA*" (or "*OBS-U*") entity [12] in a factual entity subsequence, the keyword "*no*" (or "*maybe*") is added at the beginning of the subsequence. Finally, we employ the [SEP] token to connect these subsequences, creating factual entity sequences. SEE eliminates the presentation-style elements in reports and enhances the quality of factual entity sequences, reducing the noise in the following cross-modal alignment module.

Cross-modal alignment between X-ray images and factual entity sequences. To maintain consistency in representations of the same instance across different modalities, we adopt the PRIOR [6] method to define global image-to-report alignment loss $\mathcal{L}_{global}^{R \leftarrow I}$ and global report-to-image alignment loss $\mathcal{L}_{global}^{I \leftarrow R}$. Recognizing the importance of fine-grained features in medical report generation, we utilize the PRIOR [6] method to define local image-to-report alignment loss $\mathcal{L}_{local}$. Significantly, our approach differs from the PRIOR method in that we extract text features from factual entity sequences in reports rather than the original reports. This allows our model to focus on the cross-modal alignment between medical images and factual entity sequences in reports. To summarize, the training objective for the cross-modal alignment module is $\mathcal{L}_{global}^{R \leftarrow I} + \mathcal{L}_{global}^{I \leftarrow R} + \mathcal{L}_{local}$.

2.2 Report Generation Module Based on Both Similar Historical Cases and Patient-specific Indications

A doctor formulates a particular treatment based on both patient-specific indications (i.e., the examination intention of a patient) and previous patients with similar clinical findings or symptoms. Consequently, we introduce the report generation module based on both similar historical cases and patient-specific

indications. In this Section, we will describe similar historical cases retrieval, cross-modal fusion network, and the report generation module.

Similar historical cases retrieval. Building on the pre-trained model from Section 2.1, we begin by extracting aligned image features. Given that these features implicitly preserve semantic similarity with their corresponding reference reports, we utilize a similar image matching approach (e.g., the dot product) to conduct gradient-free retrieval of similar historical cases for each sample from the training set. To enhance retrieval efficiency, we employ the Faiss tool [14] to compute the similarity between image features.

Cross-modal fusion network. To integrate information from similar historical cases, patient-specific indications, and X-ray images, we propose the cross-modal fusion network. This network comprises three Transformer Decoder layers [2,3], each featuring a self-attention sub-layer, cross-attention sub-layer, and feed-forward sub-layer. As shown in Fig. 1(b), the network automatically extracts information related to X-ray images and patient-specific indications from similar historical cases, respectively, enriching image and indication features. Subsequently, it further integrates these enriched features through a Transformer Decoder layer. Notably, even if some samples lack indications, the proposed cross-modal fusion network can fully utilize the available indication information through the output rules shown in Fig. 1(b). These processes allow the text decoder to attend to discriminative features of X-ray images, assimilate historical diagnostic information from similar cases, and understand the examination intention of patients.

Report generation module. We initialize the image and text encoders with the pre-trained model discussed in Section 2.1. Afterward, we preprocess patient-specific indications. Specifically, we delete illegal characters (e.g., "/", "_", "@") and invalid words (e.g., "*history:*", "*-year-old*", "*year old*"). When gender information is present in a patient-specific indication, we unify it as either "*man*" or "*woman*". Finally, the report generation module is optimized by minimizing the negative log-likelihood $P\left(\tilde{y}_t^i \middle| \boldsymbol{X}^i, c_K^i, I^i, \tilde{y}_{j,j<t}^i\right)$:

$$\mathcal{L}_{LM} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{t=1}^{M}\log P\left(\tilde{y}_t^i \middle| \boldsymbol{X}^i, c_K^i, I^i, \tilde{y}_{j,j<t}^i\right), \tag{1}$$

where B, M, $\boldsymbol{X}$, c_K, I, and $\tilde{y}_{j,j<t}$ denote the batch size, the maximum length of tokens generated by the text decoder, image features extracted by the image encoder, the set with K similar historical cases, the patient-specific indication, and the word sequence predicted by the text decoder for the first $t-1$ time steps, respectively.

3 Experiments

3.1 Datasets, Evaluation Metrics, and Experimental Settings

Datasets. We evaluate the effectiveness of our SEI using the MIMIC-CXR [13] dataset, following the official partitioning settings. Our approach aligns with

prior studies [5,16,20,29], utilizing the findings section of raw radiology reports as reference reports. Additionally, we filter out samples with either empty or clinically meaningless report content, such as "*Portable supine chest radiograph_ _ at 23:16 is subnitted.*". Therefore, the training, validation, and test sets include 269,239 (150,957), 2,113 (1,182), and 3,852 (2,343) chest X-ray images (reports), respectively. All reproducibility methods utilize the same test set to ensure a fair and consistent comparison.

Table 1. Comparison of our SEI with SOTA approaches on MIMIC-CXR. † means quoted results from the published literature, excluding RG and CX5, as these were not calculated in the literature. The remaining results are reproduced using the official code and checkpoints. The best values for each M_{gt} are highlighted in **bold**, with the second-best values in underlined. SEI-n denotes our SEI incorporated with information from n similar historical cases. Larger values for each metric indicate better performance.

Method	M_{gt}	NLG				CE		
		BL-2	BL-4	MTR	R_L	RG	CX5	CX14
R2Gen [5] (EMNLP'20)	100†	0.218	0.103	0.137	0.264	0.207	0.340	0.340
	Cpl.	0.209	0.097	0.135	0.266	0.211	0.339	0.338
R2GenCMN [4] (ACL'21)	100†	0.218	0.106	0.142	0.278	0.220	0.461	0.278
	Cpl.	0.198	0.090	0.133	0.268	0.223	0.464	0.393
GSKET [29] (MedIA'22)	80†	0.228	0.115	-	0.284	-	-	0.371
CGPT2 [18] (ARTMED'23)	60†	0.248	0.127	0.155	0.286	0.223	0.463	0.391
	Cpl.	0.204	0.102	0.138	0.277	0.237	0.483	0.434
M2KT [28] (MedIA'23)	80†	0.237	0.111	0.137	0.274	0.204	0.477	0.352
	Cpl.	0.204	0.085	0.133	0.244	0.210	0.483	0.413
DCL [16] (CVPR'23)	90†	-	0.109	0.150	0.284	-	-	0.373
RGRG [20] (CVPR'23)	*Cpl.*†	**0.249**	0.126	**0.168**	0.264	-	**0.547**	0.447
SEI-0 (ours)	60	**0.268**	0.146	0.164	0.300	**0.239**	0.505	0.437
	80	0.250	0.135	0.158	**0.300**	**0.250**	0.531	0.452
	90	0.244	0.131	0.156	0.299	**0.252**	0.536	0.455
	100	0.240	0.129	0.154	0.298	**0.252**	0.539	0.457
	Cpl.	0.231	0.123	0.150	**0.297**	**0.252**	0.541	0.457
SEI-1 (ours)	60	**0.268**	**0.148**	**0.167**	**0.301**	0.236	**0.509**	**0.445**
	80	**0.257**	**0.140**	**0.162**	**0.300**	0.247	**0.535**	**0.457**
	90	**0.251**	**0.137**	**0.160**	**0.300**	0.248	**0.539**	**0.459**
	100	**0.247**	**0.135**	**0.158**	**0.299**	0.249	**0.542**	**0.460**
	Cpl.	0.238	**0.128**	0.154	0.296	0.249	0.545	**0.460**

Evaluation Metrics. We utilize metrics for both conventional natural language generation (NLG) and clinical efficacy (CE) to estimate lexical similarity and clinical effectiveness between generated and reference reports. Specifically, NLG metrics include BLEU-2 (BL-2), BLEU-4 (BL-4), METEOR (MTR), and ROUGE_L (R_L), calculated using pycocoevalcap[7]. CE metrics comprise

[7] https://github.com/tylin/coco-caption

$F_{1,\text{mic-14}}$ CheXbert (CX14), $F_{1,\text{mic-5}}$ CheXbert (CX5) [19], and F_1 RadGraph (RG) [7,12], calculated by f1chexbert [19] and radgraph [12] libraries.

Experimental Settings. We regard ResNet101 [5,9] pre-trained on ImageNet as the image encoder, and a six-layer pre-trained SciBERT [1] model as the text encoder. In addition, we adopt the memory-driven Transformer, designed by R2Gen [5], as the text decoder and train it from scratch. In the first stage (i.e., training the cross-modal alignment module), we employ the AdamW optimizer with an initial learning rate of 5e-5, conducting training for 100 epochs. In the second stage (i.e., training the report generation module), we utilize the RAdam optimizer with a learning rate of 5e-5 for 30 epochs. The optimal model is selected based on cumulative scores, considering RG, CX14, and BL-4 metrics on the validation set. Subsequently, we present the results on the test set accordingly.

Table 2. Ablation studies on MIMIC-CXR in the general scenario. SEI-n represents our SEI incorporated with information from n similar historical cases. The best result is indicated in **bold**. SHC denotes similar historical cases, and w/o is without.

Settings	Model	NLG				CE		
		BL-2	**BL-4**	**MTR**	**R_L**	**RG**	**CX5**	**CX14**
(a)	Base (R2Gen [5])	0.209	0.097	0.135	0.266	0.211	0.339	0.338
(b)	(a)+cross-modal module	0.206	0.098	0.138	0.277	0.234	0.513	0.431
(c)	SEI-1 w/o indications	0.228	0.109	0.148	0.279	0.241	0.542	**0.474**
(d)	SEI-1 w/o SHC (**SEI-0**)	0.231	0.123	0.150	**0.297**	**0.252**	0.541	0.457
(e)	**SEI-1**	**0.238**	**0.128**	**0.154**	0.296	0.249	**0.545**	0.460

3.2 Main Results

To comprehensively assess coherence and integrality between generated and reference reports, we preserve the generated reports unaltered and truncate reference reports to a specific length, denoted as M_{gt}, to establish ground truth. This allows us to evaluate the model performance in different scenarios. In specific scenarios (i.e., $M_{gt} \in \{60, 80, 90, 100\}$), such as emergency diagnoses, concise medical reports prove more effective. In contrast, comprehensive and detailed reports are essential in a general scenario (i.e., $M_{gt} = Cpl.$, where *Cpl.* represents the length of complete reference reports). We compare our SEI with seven state-of-the-art (SOTA) approaches: R2Gen [5], R2GenCMN [4], GSKET [29], CvT2DistillGPT2 (CGPT2) [18], M2KT [28], DCL [16], and RGRG [20].

Results on MIMIC-CXR are presented in Table 1. Upon observation, our approach consistently outperforms previous state-of-the-art methods across diverse scenarios, achieving significant improvements on almost all metrics, notably with a notable RG score of 0.252. These experimental findings highlight the capability of our SEI to generate reports with impressive clinical efficacy and linguistic fluency in various scenarios.

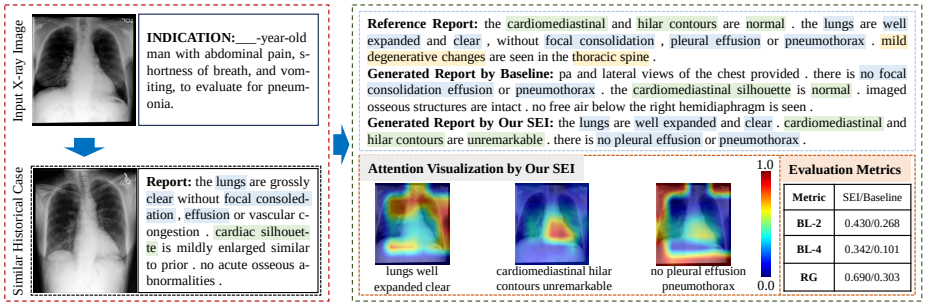

Fig. 2. An example of generated reports and attention visualization on MIMIC-CXR test set. Distinct colors in the reference report indicate the factual entity subsequence within different sentences. Generated reports and similar historical cases are highlighted in matching colors. "*Baseline*" represents the CGPT2 [18] method.

3.3 Ablation Study

Table 2 illustrates the positive effects of each component on model performance, particularly the similar historical cases and patient-specific indications. We observe from (c) and (d) that integrating them individually into the model leads to significant improvements in both NLG and CE metrics. Although SEI-1, which integrates two components simultaneously, shows performance degradation compared to (c) and (d) on certain metrics, it enhances overall performance by 4.9% and 1.9% across all metrics, respectively. This may be attributed to potential interference between these components, hindering the full exploitation of their respective strengths. In addition, the absence of indications in some samples has resulted in an unstable fusion feature space. Addressing this issue remains a topic for future research.

3.4 Qualitative Analysis

The left side of Fig. 2 presents the patient-specific indication and historical similar cases of the example, while the right side illustrates the generated reports and attention visualization. We observed that our generated report aligns with the phrase "*to evaluate for pneumonia*" in the indication section and demonstrates a high level of consistency with the reference report in terms of clinical efficacy.

4 Conclusion

In this paper, we proposed SEI for chest X-ray report generation. SEI first developed the structural entities extraction (SEE) approach to extract factual entity sequences from reports. This reduces the noise in the following cross-modal alignment module, thereby further aiding the model in gradient-free retrieval of similar historical cases from the training set. Subsequently, we proposed a cross-modal fusion network to integrate the information from X-ray images, similar

historical cases, and patient-specific indications, ensuring the text decoder attended to discriminative features of X-ray images, assimilated historical empirical information from similar cases, and understood the examination intention of patients. Experiments on MIMIC-CXR in various scenarios proved that our SEI outperformed previous state-of-the-art methods. The ablation study demonstrated the significance of SEE for the cross-modal alignment module, along with similar historical cases and patient-specific indications for the report generation module. While SEI exhibited impressive performance, it did not incorporate patient-specific indications during the cross-modal alignment phase.

Acknowledgments. The work was jointly supported by the National Science and Technology Major Project (No. 2022ZD0117103), the National Natural Science Foundations of China (No. 62272364), the provincial Key Research and Development Program of Shaanxi (No. 2024GH-ZDXM-47), the Research Project on Higher Education Teaching Reform of Shaanxi Province (No. 23JG003), and the Fundamental Research Funds for the Central Universities (No. ZYTS24090).

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

References

1. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: EMNLP. pp. 3615–3620 (2019). https://doi.org/10.18653/v1/D19-1371
2. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. In: ICCV. pp. 23346–23356 (2023). https://doi.org/10.1109/ICCV51070.2023.02139
3. Chen, Z., Li, G., Wan, X.: Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In: ACMMM. pp. 5152–5161 (2022). https://doi.org/10.1145/3503161.3547948
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: ACL. vol. 1, pp. 5904–5914 (2021). https://doi.org/10.18653/v1/2021.acl-long.459
5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: EMNLP. pp. 1439–1449 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.112
6. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., Tang, X.: Prior: Prototype representation joint learning from medical images and reports. In: ICCV. pp. 21361–21371 (2023). https://doi.org/10.1109/ICCV51070.2023.01953
7. Delbrouck, J.B., et al.: Improving the factual correctness of radiology report generation with semantic rewards. In: EMNLP. pp. 4348–4360 (2022)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. vol. 1, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
10. Hou, Z., Yan, R., Wang, Q., Lang, N., Zhou, X.: Diversity-preserving chest radiographs generation from reports in one stage. In: MICCAI. vol. 14224, pp. 482–492 (2023). https://doi.org/10.1007/978-3-031-43904-9_47

11. Huang, X., Yan, F., Xu, W., Li, M.: Multi-attention and incorporating background information model for chest x-ray image report generation. IEEE Access **7**, 154808–154817 (2019). https://doi.org/10.1109/ACCESS.2019.2947134
12. Jain, S., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. In: NeurIPS. vol. 1 (2021)
13. Johnson, A.E., et al.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
14. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019). https://doi.org/10.1109/TBDATA.2019.2921572
15. Kong, M., Huang, Z., Kuang, K., Zhu, Q., Wu, F.: Transq: Transformer-based semantic query for medical report generation. In: MICCAI. vol. 13438, pp. 610–620 (2022). https://doi.org/10.1007/978-3-031-16452-1_58
16. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: CVPR. pp. 3334–3343 (2023). https://doi.org/10.1109/CVPR52729.2023.00325
17. Nguyen, D., Chen, C., He, H., Tan, C.: Pragmatic radiology report generation. In: ML4H. vol. 225, pp. 385–402. PMLR (2023)
18. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023). https://doi.org/10.1016/j.artmed.2023.102633
19. Smit, A., et al.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: EMNLP. pp. 1500–1519 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.117
20. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR. pp. 7433–7442 (2023). https://doi.org/10.1109/CVPR52729.2023.00718
21. Tian, J., Zhong, C., Shi, Z., Xu, F.: Towards automatic diagnosis from multi-modal medical data. In: MICCAI. vol. 11797, pp. 67–74 (2019). https://doi.org/10.1007/978-3-030-33850-3_8
22. Touvron, H., et al.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023). https://doi.org/10.48550/ARXIV.2302.13971
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30 (2017)
24. Wang, F., et al.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: NeurIPS. vol. 35, pp. 33536–33549 (2022)
25. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: CVPR. pp. 11558–11567 (2023). https://doi.org/10.1109/CVPR52729.2023.01112
26. Xie, Y., Gu, L., Harada, T., Zhang, J., Xia, Y., Wu, Q.: Medim: Boost medical image representation via radiology report-guided masking. In: MICCAI. vol. 14220, pp. 13–23 (2023). https://doi.org/10.1007/978-3-031-43907-0_2
27. Yan, B., et al.: Style-aware radiology report generation with radgraph and few-shot prompting. In: EMNLP. pp. 14676–14688 (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.977
28. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023). https://doi.org/10.1016/j.media.2023.102798
29. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical Image Analysis **80**, 102510 (2022). https://doi.org/10.1016/j.media.2022.102510

30. Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: MICCAI. vol. 11769, pp. 721–729 (2019). https://doi.org/10.1007/978-3-030-32226-7_80
31. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023). https://doi.org/10.1038/s41467-023-40260-7
32. Zhang, Z., Chen, P., Sapkota, M., Yang, L.: Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references. In: MICCAI. p. 320–328 (2017). https://doi.org/10.1007/978-3-319-66179-7_37