Dia-LLaMA: Towards Large Language Model-driven CT Report Generation

Zhixuan Chen[1], Luyang Luo[2], Yequan Bie[1], and Hao Chen[1,3,4,5,6(✉)]

[1] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2] Department of Biomedical Informatics, Harvard Medical School, Boston, USA
[3] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[4] Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[5] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, The Hong Kong University of Science and Technology, Futian, Shenzhen, China
[6] State Key Laboratory of Nervous System Disorders, The Hong Kong University of Science and Technology, Hong Kong SAR, China
zchenhi@connect.ust.hk, jhc@cse.ust.hk

Abstract. Medical report generation has made notable progress, but most studies focus on chest X-rays, leaving CT report generation largely underexplored. This task poses unique challenges, including sparse diseased regions due to high-dimensional volumes, imbalanced distributions of normal and abnormal samples leading to biased predictions, and excessive template sentences that may obscure critical findings. Recently, large language models (LLMs) have demonstrated strong instruction-following capabilities, producing reliable outputs when guided by well-designed prompts, which provides a promising approach to address these issues. To this end, we propose **Dia-LLaMA**, a framework adapted from LLaMA2-7B for CT report generation with diagnostic guidance prompts. To enhance the focus on diseased areas, we introduce a disease-aware attention module to capture disease-specific information. Furthermore, we propose a disease prototype memory bank to capture common disease patterns, providing a reliable reference during diagnosis. Experiments on a large-scale chest CT report dataset demonstrated that our method outperforms previous approaches, achieving state-of-the-art results in both clinical efficacy and natural language generation metrics. The code is available at https://github.com/zhi-xuan-chen/Dia-LLaMA.

Keywords: CT Report Generation · LLM · Prototype Representation.

1 Introduction

CT report writing is a crucial part of clinical practice, offering clinicians a comprehensive summary of findings from CT volumes while effectively highlighting

✉ Corresponding author.

critical abnormalities. However, this job is tedious as it requires examining a series of CT slices and acquiring a comprehensive understanding of the CT volumes. Therefore, automated CT report generation (CTRG) is highly valuable for improving efficiency and alleviating the burden on clinicians. Inspired by the remarkable language generation capability of large language models (LLMs), several studies [9,23,1,5,7] have explored their application in report generation. However, due to the relatively limited number of CT-report pairs, previous studies have primarily focused on chest X-ray (CXR) report generation, leaving the potential of LLMs in CTRG largely unexplored. In fact, integrating LLMs into CTRG is non-trivial and presents three key challenges: 1) Abnormal areas are sparser in high-dimensional CT volumes compared to the relatively lower-dimensional CXR images, making it more difficult for models to effectively capture and interpret clinically significant regions. 2) The prevalence of normal and abnormal cases in reports may vary significantly for certain diseases [14,10]. This inherent data imbalance may cause the model to overlook infrequent abnormalities. 3) CT reports often follow a rigid template structure, with only minor modifications to describe specific abnormalities [22,12,27]. This standardized format causes models to be inclined to generate generic templates rather than accurately highlight critical abnormalities.

In this paper, we propose a novel framework that incorporates LLMs to empower CTRG, mitigating the aforementioned challenges inherent to this task. To enhance the perception of local diseased regions in CT volumes, we introduce a disease-aware attention module to effectively capture disease-level features. To improve the diagnosis of abnormalities, we propose diagnosing diseases by referencing a set of learnable disease prototypes, which encode common representations of normal and abnormal samples across different diseases. These prototypes are updated under the supervision of contrastive loss to ensure distinctiveness between normal and abnormal samples, providing a valuable reference for disease diagnosis, especially for abnormalities that rarely appear in the dataset. Furthermore, we propose emphasizing critical disease information by embedding it into prompts for LLM. Leveraging the powerful instruction-following capability of LLMs, this method enables the generation of coherent and comprehensive reports that effectively highlight significant abnormalities. Experiments conducted on a large-scale publicly available chest CT report dataset demonstrate that our proposed framework achieves state-of-the-art (SOTA) performance in both clinical efficacy (CE) and natural language generation (NLG) metrics.

2 Method

2.1 Framework

The overall architecture is shown in Figure 1. To integrate LLMs for report generation, we employ a structured prompt that combines visual embeddings with key diagnostic information. Our designed prompt comprises two segments: $\mathcal{P} = \{\mathcal{S}, \mathcal{D}\}$, where the first segment $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ consists of special tokens s_n for visual embeddings and the second segment $\mathcal{D} = \{d_1, d_2, \ldots, d_L\}$

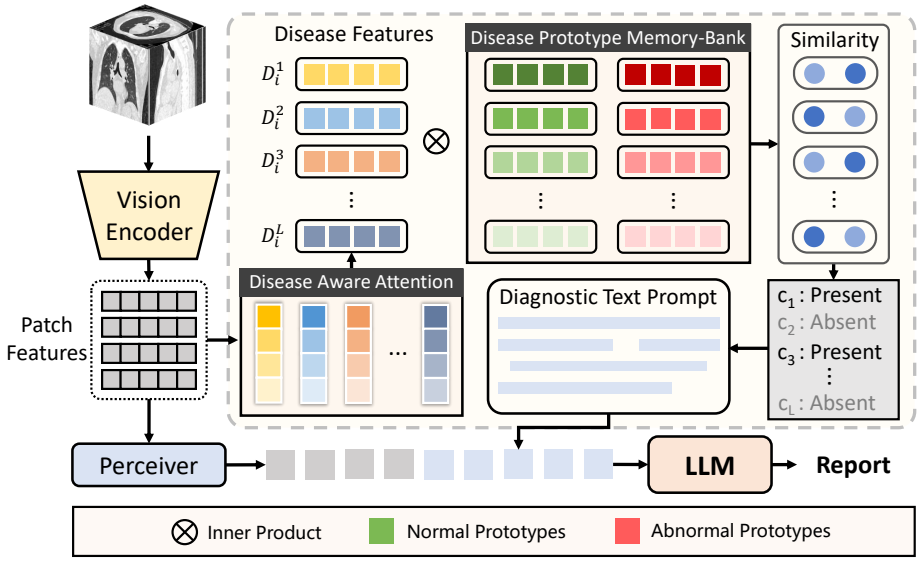

Fig. 1. The overall architecture. Disease-aware attention is employed to extract disease features, which in turn update disease prototypes that capture common representations across various diseases. Diagnosis results are generated by comparing disease features with the prototypes. The resulting diagnostic information, along with the visual features, is then converted into prompts for the LLM to generate reports.

represents diagnostic text prompt tokens d_l for the l_{th} disease. The N and L denote the number of s_n and d_l, respectively. Let $\mathcal{R} = \{r_1, r_2, \ldots, r_T\}$ denotes a generated report, where r_t represents the token at timestep t and T is the length of the report. The generation process of the LLM f_l is formulated as follows:

$$r_t = f_l(\mathcal{P}, \mathcal{R}^-) = f_l(\mathcal{S}, \mathcal{D}, \mathcal{R}^-) = f_l(s_1, \ldots, s_N, d_1, \ldots, d_L, r_1, \ldots, r_{t-1}), \quad (1)$$

where $\mathcal{R}^-$ represents the generated report at timestep $t-1$. This process is optimized by minimizing the language modeling loss $\mathcal{L}_{LM}$:

$$\mathcal{L}_{LM} = -\sum_{t=1}^{T} \log p(r_t | s_1, \ldots, s_N, d_1, \ldots, d_L, r_1, \ldots, r_{t-1}). \quad (2)$$

To extract visual embeddings, a volume encoder f_v is employed to encode the i_{th} CT volume V_i into patch features, which are then projected into the LLM's embedding space by a perceiver f_p:

$$f_v(V_i) = A_i = \{A_i^1, A_i^2, \ldots, A_i^M\}, \quad (3)$$

$$f_p(A_i) = X_i = \{X_i^1, X_i^2, \ldots, X_i^N\}, \quad (4)$$

where $A_i^m \in \mathbb{R}^c$ represents a patch feature, $X_i^n \in \mathbb{R}^d$ represents visual embedding, c and d denote the feature and embedding dimension, M and N represent the number of patch features and visual embeddings, respectively.

To derive diagnostic information, we first apply disease-aware attention (Section 2.2) to extract disease-level features D_i from patch features A_i. To provide a reference during disease diagnosis, we construct a disease prototype memory bank (Section 2.3) to capture common representations of various diseases. The diagnostic results are obtained by comparing disease features with prototypes and then converted into diagnostic text prompts (Section 2.4) for the LLM.

2.2 Disease-Aware Attention

The previous work [10] utilized pooled patch features for disease diagnosis, which may result in unreliable outcomes due to the mixed disease information. To alleviate this issue, we propose a disease-aware attention (DAA) module to extract disease-level features from patch features. Specifically, we assign a set of learnable attention weights to each disease. The patch features from the vision encoder f_v are element-wise multiplied by attention weights and subsequently aggregated to obtain the disease-level features. This process can be formulated as:

$$D_i = \sum_{m=1}^{M} (\text{Softmax}(\mathbf{W}_D) \otimes A_i)_m, \tag{5}$$

where $D_i \in \mathbb{R}^{L \times c}$ represents the aggregated disease features, $\mathbf{W}_D \in \mathbb{R}^{L \times M \times 1}$ denotes the disease-aware attention weights and $A_i \in \mathbb{R}^{1 \times M \times c}$ encapsulates the patch features. The operator $\otimes$ represents element-wise multiplication with automatic broadcasting [19]. The resulting disease features D_i are then utilized for disease diagnosis.

2.3 Disease Prototype Memory Bank

To improve the diagnostic accuracy for infrequent abnormalities, we introduce a disease prototype memory bank (DPM) as a reference during diagnosis. The diagnostic results are obtained by comparing the similarity between disease-level features and a set of learnable prototypes. Specifically, the DPM includes both abnormal prototypes $\mathbf{P}_1^l$ and normal prototypes $\mathbf{P}_0^l$ to represent the characteristic features of each disease in its presence and absence, respectively. The prototypes are initialized from a standard normal distribution $\mathcal{N}(0, 1)$ and updated via a contrastive loss [17], which pulls the positive pairs closer and pushes the negative pairs farther. For the disease features D_i^l in our method, the positive case $\mathbf{P}_{y_i^l}^l$ and negative case $\mathbf{P}_{1-y_i^l}^l$ are determined based on the disease label y_i^l. The contrastive disease-prototype loss $\mathcal{L}_{DP}$ is defined as

$$\mathcal{L}_{DP} = -\frac{1}{BL} \sum_{i=1}^{B} \sum_{l=1}^{L} \log \frac{\exp(D_i^l \cdot \mathbf{P}_{y_i^l}^l / \tau)}{\exp(D_i^l \cdot \mathbf{P}_{y_i^l}^l / \tau) + \exp(D_i^l \cdot \mathbf{P}_{1-y_i^l}^l / \tau)}, \tag{6}$$

where y_i^l denotes the label of the l_{th} disease, and τ is the temperature parameter.

2.4 Diagnostic Text Prompts

The critical abnormality information is essential in medical reports [22]. Despite the strong capabilities of LLM, directly capturing abnormalities from visual embeddings without additional guidance is still challenging, which is validated in Section 3.3. Therefore, we introduce diagnostic text prompts (DTP), leveraging the diagnostic results as instruction for LLM. Specifically, the diagnostic results are converted into text prompts $\mathcal{D}$, which follows a template description "*The* {*disease name*} *is* [*disease state*]". For instance, the diagnostic result c_1: *Present* in Figure 1 is interpreted as *The cardiomegaly is present in this image*, where c_1 represents the *cardiomegaly* disease.

The overall loss in our framework is defined as the sum of the disease-prototype loss $\mathcal{L}_{DP}$ and the language modeling loss $\mathcal{L}_{LM}$:

$$\mathcal{L} = \mathcal{L}_{DP} + \mathcal{L}_{LM}. \tag{7}$$

3 Experiments and Results

3.1 Datasets and Metrics

We adopted a large-scale CT report dataset (CTRG-Chest-548K [22]) to evaluate our method and the compared methods. This dataset comprises 1,804 CT-report pairs, with 80% of the data used for training and 20% for testing. Following previous works [10,26], we employ the pre-trained report labeler CheXbert [21] to extract disease labels. Despite being pre-trained on the CXR dataset (MIMIC [11]), CheXbert remains effective in our experiments due to the similarity in content between chest CT and CXR reports. While it originally identifies 14 diseases, some are too rare in our CT reports, so we selected 8 diseases to be included in the diagnostic prompt.

For evaluation, both NLG and CE metrics are adopted. NLG metrics include BLEU [18], METEOR [6], and ROUGE-L [13]. Following the CE metrics setting in [16,10], we assess Precision, Recall, and F1 score with CheXbert [21].

3.2 Implementation details

For comparison, we evaluated our model against the CT report generation methods SL-DG [22] and RadFM [25]. To ensure a fair comparison, we aligned the LLM in RadFM with the one used in our experiments and fine-tuned it on the CT report dataset. Given the limited research on CT report generation (CTRG), we also compared our method with state-of-the-art (SOTA) approaches in chest X-ray (CXR) report generation, including R2Gen [4], R2GenCMN [3], M2KT [26], and PromptMRG [10]. These methods support multi-image input, allowing them to treat the slices within a CT volume as multiple images for CT report generation. We adopt a pre-trained ViT3D [25] as our volume encoder, with each input volume resized to $256 \times 256 \times 64$. This encoder employs a 3D patch extraction strategy, partitioning the volume into multiple 3D patches, each of which

Table 1. The performance of our model compared with other SOTA methods on the CTRG-Chest-548K [22] dataset. $*$ indicates results cited from the original paper. Our method is highlighted in green. The best results and the second-best results are highlighted in **bold** and underlined, respectively.

METHOD	CE Metrics			NLG Metrcis			
	Pre.	Rec.	F1	BL-1	BL-4	MTR	RG-L
R2Gen [4]	0.207	0.121	0.144	34.11	23.39	21.40	**47.75**
R2GenCMN [3]	0.158	0.100	0.114	35.88	23.37	21.43	45.94
M2KT [26]	0.220	0.119	0.145	46.09	21.93	25.20	36.47
PromptMRG [10]	0.290	0.330	0.290	47.73	23.02	22.87	37.35
SL-DG* [22]	-	-	-	-	23.70	21.90	43.80
RadFM [25]	0.403	0.361	0.345	46.70	24.70	24.01	38.98
Ours	**0.421**	**0.387**	**0.372**	**51.16**	**29.64**	**26.28**	42.15

is embedded into a corresponding patch feature. We employ LLaMA2-7B [24] as the LLM in all our experiments and adopt LoRA [8] for parameter-efficient fine-tuning. In our configuration, LoRA is applied with a rank of 8, a scaling factor (`lora_alpha`) of 32, and a dropout rate of 0.1. During training, we utilized AdamW [15] as the optimizer, with an initial learning rate of 5e-5, following a constant learning rate schedule that includes a warmup phase. The model was trained on two RTX 3090 GPUs for about 16 hours, built with PyTorch 2.0. The training involved 20 epochs, with an effective batch size of 16. To optimize memory usage, we employed the ZeRO [20] stage 2 training strategy in conjunction with gradient checkpointing [2].

3.3 Comparison and Analysis

The Table 1 shows the comparison results on CTRG-Chest-548K [22] dataset. We observed that the proposed method achieves SOTA performance across all CE metrics and the majority of NLG metrics. For CE metrics, our model surpassed the second-best method by 4.5%, 7.2% and 7.8% in precision, recall, and F1 score, respectively. This demonstrates the superiority of our model in generating reports with higher diagnostic accuracy. In terms of NLG metrics, our method also achieved SOTA performance. Regarding the BLEU-1, BLEU-4, and METEOR metrics, our approach obtained improvements of 7.2%, 20%, and 4.3%, respectively, compared to the inferior methods. The relatively lower ROUGE-L score could be attributed to the inherent nature of the metric, which assesses reports based solely on sentence matching without considering semantic similarity. Therefore, methods that leverage memory mechanisms [4,3] can easily achieve higher ROUGE-L scores by generating common template sentences. However, for metrics like METEOR, which consider semantic relevance, our method outperforms other methods by a significant margin, demonstrating that the reports generated by our method exhibit superior quality.

Table 2. Ablation study of each module on CTRG-Chest-548K [22] dataset.

DPM	DAA	DTP	CE Metrics			NLG Metrcis			
			Pre.	Rec.	F1	BL-1	BL-4	MTR	RG-L
✗	✗	✗	0.403	0.361	0.345	46.70	24.70	24.01	38.98
✗	✗	✓	0.415	0.336	0.347	45.74	27.05	24.80	42.29
✗	✓	✓	0.424	0.347	0.358	44.22	26.38	24.34	42.68
✓	✗	✓	**0.437**	0.313	0.339	44.06	27.10	24.46	**44.5**
✓	✓	✓	0.421	**0.387**	**0.372**	**51.16**	**29.64**	**26.28**	42.15

Ablation Study To demonstrate the effectiveness of all the proposed components, we conducted a thorough ablation study, as shown in Table 2. We adopted RadFM [25] as the baseline, which lacks additional diagnostic information. For the method that solely incorporates DTP, we directly input the pooled patch features into a classification head to generate diagnostic prompts. We can see its improvements in almost all metrics compared to the baseline, which confirms the significance of incorporating diagnostic information for guiding LLM in report generation. When the DAA is incorporated, the CE metrics show further improvement, validating the significance of emphasizing disease information in the volume features. After integrating the DPM, our complete method with all proposed components achieved SOTA performance in most metrics. We also assessed the method without DAA alone, which resulted in a subpar F1 score, underscoring the essential role of fine-grained disease features in diagnosis. A representative qualitative example is presented in Figure 3. It demonstrates that our method captures more critical abnormality information compared to the baseline and achieves higher diagnostic accuracy.

We also assessed the F1 scores for each disease separately to validate the diagnostic performance of our method across diseases, as presented in Figure 2. The diseases are arranged in ascending order based on the number of abnormal samples, with the final group representing the average F1 score across all 8 diseases. We observed that the method using only DTP performed poorly when the number of abnormal samples was limited compared to the normal samples. This demonstrates that diagnosis based on the classification head can be easily affected by data imbalance due to the lack of reference during diagnosis. In contrast, our complete method with DAA and DPM achieved a significantly higher F1 score, particularly for diseases with fewer abnormal samples. This validates that our proposed method can address the challenge of data imbalance by emphasizing disease features and providing diagnostic reference with learned disease prototypes, thereby improving overall diagnostic accuracy.

Moreover, we conducted an ablation study on different prompt types to find the appropriate one, as presented in Table 3. Specifically, the *None* prompt indicates that no diagnostic result is used as the prompt. The *Text* prompt is just the DTP proposed in Section 2.4, while *Token* prompt indicates we incorporated additional special tokens $<POS\text{-}l>$ and $<NEG\text{-}l>$ to represent the disease diagnosis instead of text tokens. For the *Feature* prompt, we directly utilize the

Table 3. The comparison of different prompt types. *None* represents the baseline with only visual embedding as the prompt. *Text* represents the diagnostic textual prompt. *Token* represents the special token prompt, while *Feature* represents the disease prototype prompt.

Prompt	B-4	Pre.	Rec.	F1
None	24.70	0.403	0.361	0.345
Text	29.64	0.421	0.387	0.372
Token	25.40	0.363	0.387	0.340
Feature	23.10	0.327	0.359	0.310

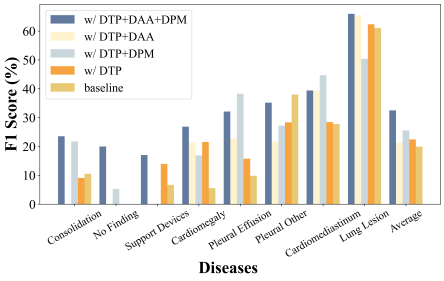

Fig. 2. Comparison of the F1 score (%) of each disease across different settings. The diseases are sorted in ascending order based on the number of abnormal samples in the dataset.

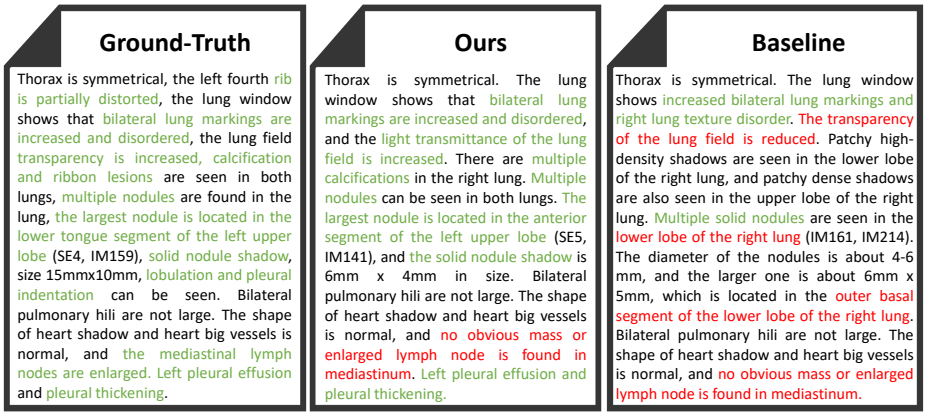

Fig. 3. Qualitative example of the Baseline and our method. Green indicates the consistent abnormality information, while Red represents the incorrect content.

disease prototypes $\mathbf{P}_1^l$ or $\mathbf{P}_0^l$ as prompt features, selected based on the predicted diagnostic results. These features are then fed into the LLM, along with visual features, to generate reports. The results indicate that the *Text* prompt obtained the most significant enhancement relative to the *None* prompt, so we adopted text prompts as the default prompt type.

4 Conclusion

In this work, we propose a novel CTRG framework called **Dia-LLaMA**, which effectively adapts LLaMA2-7B [24] to generate reports with diagnostic guidance prompts. Specifically, we adopt a disease-aware attention module to obtain disease-level features, enabling fine-grained perception tailored to different diseases. Additionally, a disease prototype memory bank is proposed to capture

common representations of various diseases. The diagnosis results are obtained by comparing disease features with their corresponding prototypes, effectively mitigating the negative impacts of data imbalance by providing a reliable diagnostic reference. We then interpret the diagnosis results into textual prompts as critical instruction for LLM to generate reports, achieving both linguistic coherency and outstanding diagnostic performance. Experiments on the CTRG-Chest-548K [22] dataset demonstrated the superiority of our method over other SOTA methods. We acknowledge the limitation of the current work, which focuses solely on CT report generation. In future work, we will continue to explore the potential of LLMs and develop a framework capable of generating reports across all radiology modalities.

Acknowledgments. This work was supported by the Hong Kong Innovation and Technology Commission (Project No. GHP/006/22GD and ITCPD/17-9), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-401/22-N).

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

References

1. Che, H., Jin, H., Guo, Z., Lin, Y., Jin, C., Chen, H.: Llm-driven medical report generation via communication-efficient heterogeneous federated learning. arXiv preprint arXiv:2506.17562 (2025)
2. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
5. Chen, Z., Bie, Y., Jin, H., Chen, H.: Large language model with region-guided referring and grounding for ct report generation. IEEE Transactions on Medical Imaging (2025)
6. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on statistical machine translation. pp. 85–91 (2011)
7. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv e-prints pp. arXiv–2404 (2024)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
9. Hyland, S.L., Bannur, S., Bouzid, K., Castro, D.C., Ranjit, M., Schwaighofer, A., Pérez-García, F., Salvatelli, V., Srivastav, S., Thieme, A., et al.: Maira-1: A specialised large multimodal model for radiology report generation. arXiv preprint arXiv:2311.13668 (2023)

10. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2607–2615 (2024)
11. 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)
12. 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)
13. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
14. Liu, G., Liao, Y., Wang, F., Zhang, B., Zhang, L., Liang, X., Wan, X., Li, S., Li, Z., Zhang, S., et al.: Medical-vlbert: Medical visual language bert for covid-19 ct report generation with alternate learning. IEEE Transactions on Neural Networks and Learning Systems **32**(9), 3786–3797 (2021)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. Artificial intelligence in medicine **144**, 102633 (2023)
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. 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)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. pp. 1–16. IEEE (2020)
21. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020)
22. Tang, Y., Yang, H., Zhang, L., Yuan, Y.: Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. Expert Systems with Applications **237**, 121442 (2024)
23. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)
24. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
25. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023)
26. 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)

27. Yang, S., Ji, J., Zhang, X., Liu, Y., Wang, Z.: Weakly guided hierarchical encoder-decoder network for brain ct report generation. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 568–573. IEEE (2021)