

KARGEN: Knowledge-enhanced Automated Radiology Report Generation Using Large Language Models

Yingshu Li¹, Zhanyu Wang¹, Yunyi Liu¹, Lei Wang², Lingqiao Liu³, and Luping Zhou¹(✉)

¹ University of Sydney, Sydney, NSW, Australia
{yingshu.li, zhanyu.wang, yunyi.liu1, luping.zhou}@sydney.edu.au

² University of Wollongong, Wollongong, NSW, Australia
leiw@uow.edu.au

³ The University of Adelaide, Australia
lingqiao.liu@adelaide.edu.au

Abstract. Harnessing the robust capabilities of Large Language Models (LLMs) for narrative generation, logical reasoning, and common-sense knowledge integration, this study delves into utilizing LLMs to enhance automated radiology report generation (R2Gen). Despite the wealth of knowledge within LLMs, efficiently triggering relevant knowledge within these large models for specific tasks like R2Gen poses a critical research challenge. This paper presents KARGEN, a **K**nowledge-enhanced **A**utomated radiology **R**eport **G**ENERation framework based on LLMs. Utilizing a frozen LLM to generate reports, the framework integrates a knowledge graph to unlock chest disease-related knowledge within the LLM to enhance the clinical utility of generated reports. This is achieved by leveraging the knowledge graph to distill disease-related features in a designed way. Since a radiology report encompasses both normal and disease-related findings, the extracted graph-enhanced disease-related features are integrated with regional image features, attending to both aspects. We explore two fusion methods to automatically prioritize and select the most relevant features. The fused features are employed by LLM to generate reports that are more sensitive to diseases and of improved quality. Our approach demonstrates promising results on the MIMIC-CXR and IU-Xray datasets. Our code will be available on GitHub.

Keywords: Radiology Report Generation · Medical Domain Knowledge Graph · Large Language Models.

1 Introduction

Automated radiology report generation (R2Gen) is gaining traction due to its potential to streamline the time-consuming and error-prone task of medical image reading and report writing. Unlike generic image captioning tasks [25,13,4],

✉Corresponding Author

which focus on concise summaries of image contents, R2Gen involves generating detailed paragraphs covering both normal and pathological findings in radiology images. Various approaches address this challenge [24,3,2]. For instance, hierarchically structured LSTM [27,24] and memory-driven modules [3,2] enhance long-term memory capabilities. Data deviation, where normal contents dominate, is another challenge [20,11]. Efforts to tackle this involve improving image-text attention, aligning features, and incorporating external domain knowledge [24,23,11]. Some studies [23] leverage additional disease classification tasks, while others [11,26] utilize knowledge graphs to capture disease-related information based on medical domain knowledge.

In the past two years, large language models (LLMs) [18] have demonstrated significant capabilities in generating more human-like, coherent, and contextually relevant responses, utilizing their extensive knowledge base. This potential has also been explored to combat the aforementioned challenges for R2Gen [22,17]. However, despite the wealth of knowledge within LLMs, efficiently triggering relevant knowledge within these large models for specific tasks like R2Gen could pose a critical research challenge. Current methods, relying primarily on visual prompts from regional image features, may struggle to capture detailed, disease-related information to effectively prompt LLMs for R2Gen. Although [17] trained a disease classifier and constructed its output as an additional text prompt, the information provided remains arguably sparse as clues for diseases.

In this paper, we present KARGEN, a novel Knowledge-enhanced Automated radiology Report GENERation framework based on LLMs. To the best of our knowledge, this is the first exploration of integrating a disease-specific knowledge graph to activate and unlock pertinent medical domain knowledge within LLMs. Diverging from previous approaches that constructed graph convolutional networks (GCNs) solely based on image or text features, our method integrates both text and image features to define graph nodes, linking regional image features with the text embedding of disease classes. Our approach fosters a comprehensive fusion of inter-disease features, allowing us to capture fine-grained disease-related features and interrelationships among diseases. Moreover, unlike prior methods that merely use graph-enhanced features for R2Gen, we advocate for the integration of both graph-enhanced disease-related features and regional image features to attend to both normal and disease-related findings in a radiology report. We therefore develop two fusion methods, operating at either individual feature element or modality (feature types) level, to effectively prioritize the most relevant features. These fused features are then leveraged to prompt LLMs to generate reports to become more sensitive to diseases and achieve improved quality.

Our main contributions are summarized as follows:

- (1) We present a novel framework that integrates a medical domain knowledge graph with LLMs for R2Gen. It demonstrates, for the first time, that despite the wealth of knowledge within LLMs, the incorporation of a specific knowledge graph encoding disease information is necessary and beneficial for activating relevant knowledge in LLMs for R2Gen.

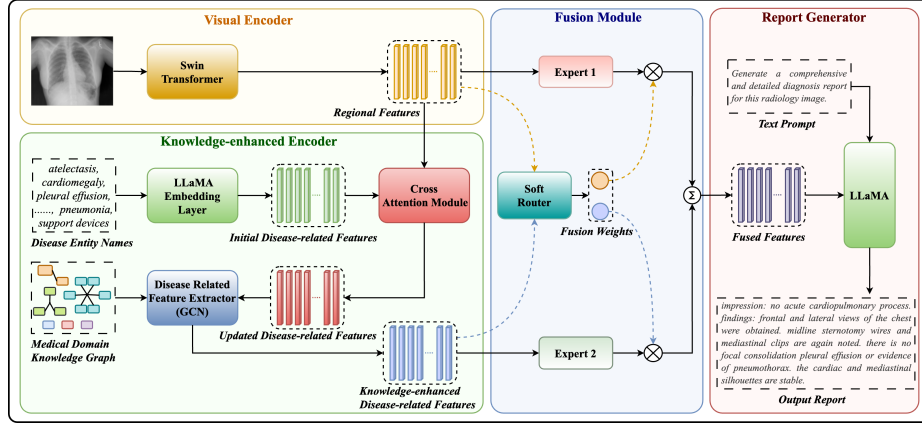


Fig. 1. An overview of the KARGEN framework, which comprises a visual encoder, a knowledge-enhanced encoder, a fusion module and a report generator.

- (2) Our model includes a novel knowledge graph for extracting disease-related features, along with two alternative strategies for a feature fusion component. These strategies effectively integrate graph-enhanced disease-related features and regional image features, enabling the model to attend to both normal and disease-related content in the generated reports.
- (3) Our method, validated on two public datasets IU-Xray and MIMIC-CXR, outperforms multiple relevant state-of-the-art methods on various evaluation metrics including the most recent clinic-related ones.

2 Methodology

Our framework consists of four main components: a visual feature encoder, a knowledge-enhanced feature encoder, a feature fusion module, and a report generator. The visual feature encoder extracts regional features from a chest x-ray image, which are then processed by the knowledge-enhanced feature encoder to ‘distill’ disease-related information guided by a medical knowledge graph. The resulting knowledge-enhanced disease-related features are fused with the regional image features in the feature fusion module and used to prompt the LLaMA-based report generator for R2Gen. Fig. 1 gives an overview of KARGEN.

2.1 Feature Extraction

Regional Feature Extraction Given an input X-ray image $\mathbf{X}_v$, we initially extract regional image features $\mathbf{Z}_v = \text{Swin}(\mathbf{X}_v; \theta_v)$, utilizing a pre-trained Swin Transformer [12], where $\mathbf{Z}_v \in \mathbb{R}^{S \times d_v}$ (S : the number of features; d_v : the dimensionality of each feature; θ_v : the parameters of the Swin Transformer).

Medical Domain Knowledge Graph Focusing on disease-related features in medical imaging, especially for interrelated chest diseases, is critical. We propose a medical domain knowledge graph to extract chest disease features, incorporating 14 terms from the Chexpert [6]. Each disease entity is represented by the word embedding of its name, obtained using the LLaMA Word Embedding Layer. The connections are illustrated in Fig. 2, highlighting that abnormalities within the same region exhibit stronger correlations than those across different organs. This guides our analysis of diseases in the lungs, heart, and pleura in chest X-ray images [30], capturing nuanced relationships effectively.

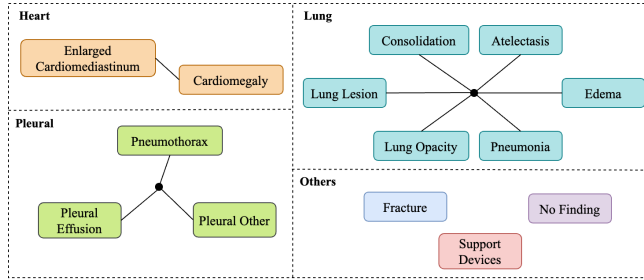


Fig. 2. The medical domain knowledge graph is constructed based on the correlations among various diseases, where diseases that are linked together are interconnected.

Diseased-related Feature Extraction Based on the knowledge graph, we construct a GCN to aggregate disease-related features. Our GCN comprises three layers. At each layer l ($l = 1, 2$ and 3), there are two primary phases: 1) the propagation of information throughout the graph, and 2) the updating of disease-related features. Let $\mathbf{N}^l$ denote the node features in the l -th layer, $\mathbf{A}$ denote the adjacency matrix governed by the knowledge graph, and $\mathbf{A}' = \mathbf{D}^{-1/2} \cdot \mathbf{A} \cdot \mathbf{D}^{-1/2}$ ($\mathbf{D}$ is the degree matrix of $\mathbf{A}$). The entire process can be formulated as

$$\begin{aligned} \mathbf{N}_{phase1}^l &= GELU(LN((\mathbf{N}^{l-1} \cdot \mathbf{W}^l) \mathbf{A}')), \\ \mathbf{N}^l &= GELU(LN((\mathbf{N}^{l-1} + \mathbf{N}_{phase1}^l) \cdot \mathbf{W}_{update}^l + \mathbf{N}^{l-1})). \end{aligned} \quad (1)$$

Here $\mathbf{W}^l$ and $\mathbf{W}_{update}^l$ are learnable parameters for information propagation and updating. $LN(\cdot)$ denotes variants of layer normalization.

The initial node features $\mathbf{N}^0$ are determined by using the disease entity name embeddings $\mathbf{E} \in \mathbb{R}^{M \times d_w} = [\mathbf{e}_1, \dots, \mathbf{e}_i, \dots, \mathbf{e}_M]^T$ ($M = 14$) to query the regional image features $\mathbf{Z}_v$ output by the visual encoder via multi-head attention:

$$\begin{aligned} \mathbf{N}^0 &= MHA(\mathbf{E}, \mathbf{Z}_v) = Concat(head_1, \dots, head_h) \mathbf{W}^O, \\ head_h &= Softmax\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_k}}\right) \mathbf{V}_h, \\ \mathbf{Q}_h &= \mathbf{E} \mathbf{W}_h^Q, \quad \mathbf{K}_h = \mathbf{Z}_v \mathbf{W}_h^K, \quad \mathbf{V}_h = \mathbf{Z}_v \mathbf{W}_h^V, \end{aligned} \quad (2)$$

where $\mathbf{W}_h^Q$, $\mathbf{W}_h^K$, $\mathbf{W}_h^V$, and $\mathbf{W}^O$ are learnable parameters.

To summarize the above process, the regional image features $\mathbf{Z}_v$ output by the visual encoder and the disease entity name embeddings $\mathbf{E}$ are cross-attended and passed through a three-layer GCN whose adjacency matrix is defined by our knowledge graph to output the disease-related feature $\mathbf{Z}_g = \mathbf{N}^3$.

2.2 Feature Fusion

After obtaining the regional features $\mathbf{Z}_v \in \mathbb{R}^{S \times d_v}$, and the knowledge-enhanced disease-related features $\mathbf{Z}_g \in \mathbb{R}^{M \times d_v}$, we employed a multi-head attention network to align their dimensions by using $\mathbf{Z}_v$, $\mathbf{Z}_g$ and $\mathbf{Z}_g$ as the query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$). The attention output, denoted as $\tilde{\mathbf{Z}}_g \in \mathbb{R}^{S \times d_v}$, shares the same dimensions as $\mathbf{Z}_v$. In the following, we propose two fusion strategies designed to integrate these two types of features.

Element-wise Fusion This approach uses an element-wise weighted sum for the final integrated feature representation, employing a trainable gate to determine the importance of each element in the two feature types. The fused features $\mathbf{Z}_f \in \mathbb{R}^{S \times d_v}$ is obtained by:

$$\begin{aligned} \mathbf{Z}_f &= \mathbf{gate} \odot \mathbf{Z}_v + (1 - \mathbf{gate}) \odot \tilde{\mathbf{Z}}_g, \\ \mathbf{gate} &= \text{sigmoid}([\mathbf{Z}_v; \tilde{\mathbf{Z}}_g] \cdot \mathbf{W}^g), \end{aligned} \quad (3)$$

where $[\mathbf{Z}_v; \tilde{\mathbf{Z}}_g]$ represents the concatenation of $\mathbf{Z}_v$ and $\tilde{\mathbf{Z}}_g$, and $\mathbf{W}^g$ is a learnable parameter. The operation $\odot$ signifies element-wise multiplication.

Modality-wise Fusion Inspired by the Mixture of experts (ME) [14], we designed two distinct expert networks: one to process disease-related features $\tilde{\mathbf{Z}}_g$ and the other for general regional features $\mathbf{Z}_v$. To dynamically allocate the contribution of each expert’s output, we put forward a soft router module, represented by $\mathbf{G}(x)$, functioning as a gating network. This gate is implemented as a multi-layer perceptron (MLP). Unlike the element-wise fusion operating at the individual element level, modality-wise fusion treats each feature set as an integral unit for combination. The combined output is formulated as:

$$\mathbf{Z}_f = g_1 E_1(\mathbf{Z}_v) + g_2 E_2(\tilde{\mathbf{Z}}_g), \quad [g_1, g_2] = \mathbf{G}(\mathbf{Z}_v, \tilde{\mathbf{Z}}_g). \quad (4)$$

Here, E_1 and E_2 denote the expert networks comprising of a linear layer and layer normalization. The soft router $\mathbf{G}(\mathbf{x})$ assesses and decides the relevance of each expert (E_1 and E_2) in the fusion process. The weights g_1 and g_2 are computed as probability values, indicating the importance assigned to each expert’s output in the final feature representation $\mathbf{Z}_f$.

2.3 Report generation

We employ LLaMA2-7B to generate radiology reports, leveraging the fused features as the visual prompt. Our instruction prompt is designed following the

template of LLaMA2. Given a set of fused features $\mathbf{Z}_f$ output by the feature fusion module according to Eqn. 4, our prompt is designed as: $[/INST] \mathbf{Z}_t <feats> \mathbf{Z}_f </feats> [/INST].$, where $\mathbf{Z}_t$ is a constant instruction text: *‘Generate a comprehensive and detailed diagnosis report for this radiology image.’*. Before input to LLaMA2, all text words in the prompt are tokenized and embedded by LLaMA’s tokenizer and word embedding layers. Recall that $\mathbf{X}_v$ denotes the input image. Our overall model is optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{CE}(\theta) = - \sum_{i=1}^{N_r} \log p_{\theta}(t_i^* | \mathbf{X}_v, \mathbf{Z}_t, t_{1:i-1}^*), \quad (5)$$

where θ denotes model parameters, and t_i^* is the i -th word in the ground truth report with a length of N_r words.

3 Experiments

Datasets Two widely used benchmarks are involved in our experiments. IU-Xray is from Indiana University Chest X-ray Collection (IU-Xray) [5], comprising 3,955 radiology reports linked to 7,470 chest X-ray images. Following the partitioning guidelines of [3], we divided the dataset into training, testing, and validation sets with a ratio of 7:1:2.

MIMIC-CXR [8] comprises 377,110 chest X-ray images and 227,835 associated reports from 64,588 patients at the Beth Israel Deaconess Medical Center (2011-2016). For consistency and fair comparison, we utilized the dataset’s division defined by [3], i.e., 270790 images for training and 3858 for testing.

Implementation Details In this work, we employed LLaMA2-7B¹ as the LLM and Swin Transformer² as the visual encoder. We used 3 layers GCN to aggregate the disease-related features through the medical domain knowledge graph. The model was trained on two NVIDIA A6000 48GB GPUs, utilizing a mini-batch size of 8 and a learning rate of 1e-4. For testing, a beam search strategy was adopted with a beam width of 3 to balance between computational efficiency and output quality.

Evaluation Metrics We used traditional natural language generation (NLG) metrics (e.g., BLEU [16], ROUGE-L [10], METEOR [1], and CIDEr [19]), as well as recent clinic-related metrics RadGraph F1 [7] and BERTScore [29], following insights from [28]. The latter offers a closer alignment with radiologist assessments than NLG metrics and the Chexpert [6] clinical efficacy score [28]. Moreover, we incorporated the RadCliQ metric [28], a comprehensive measure that combines individual metrics to better correlate with radiologist evaluations³.

¹ <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

² <https://huggingface.co/microsoft/swin-base-patch4-window7-224>

³ <https://github.com/rajpurkarlab/CXR-Report-Metric/tree/v1.1.0>

Table 1. Comparison on MIMIC-CXR and IU-Xray datasets. The highest scores are highlighted in bold, the second-highest scores are indicated with an underline.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
MIMIC-CXR	Show-Tell [25]	0.308	0.190	0.125	0.088	0.256	0.122	0.096
	AdaAtt [13]	0.314	0.198	0.132	0.094	0.267	0.128	0.131
	M2Transformer [4]	0.332	0.210	0.142	0.101	0.264	0.134	0.142
	R2Gen [†] [3]	0.353	0.218	0.145	0.103	0.277	0.142	-
	R2GenCMN [†] [2]	0.353	0.218	0.148	0.106	0.278	0.142	-
	PPKED [†] [11]	0.36	0.224	0.149	0.106	0.284	0.149	0.237
	GSK [†] [26]	0.363	0.228	0.156	0.115	0.284	-	0.203
	MSAT [†] [23]	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	METransformer [†] [21]	0.386	0.250	0.169	0.124	0.291	0.152	<u>0.362</u>
	CvT2DistilGPT2 [†] [15]	0.393	0.248	0.171	0.127	-	0.155	0.389
	RaDialog-RG [†] [17]	0.346	-	-	0.095	0.271	0.140	-
	R2GenGPT [†] [22]	<u>0.411</u>	<u>0.267</u>	<u>0.186</u>	<u>0.134</u>	<u>0.297</u>	<u>0.160</u>	0.269
	Ours (Modality-wise Fusion)	0.417	0.274	0.192	0.140	0.305	0.165	0.289
IU-Xray	Show-Tell [25]	0.243	0.130	0.108	0.078	0.307	0.157	0.197
	AdaAtt [13]	0.284	0.207	0.150	0.126	0.311	0.165	0.268
	M2transformer [4]	0.402	0.284	0.168	0.143	0.328	0.170	0.332
	R2Gen [†] [3]	0.470	0.304	0.219	0.165	0.371	0.187	-
	R2GenCMN [†] [2]	0.475	0.309	0.222	0.170	0.375	0.191	-
	KERP [†] [9]	0.482	0.325	0.226	0.162	0.339	-	0.280
	PPKED [†] [11]	0.483	0.315	0.224	0.168	0.376	0.190	0.351
	MSAT [†] [23]	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	METransformer [†] [21]	0.483	<u>0.322</u>	<u>0.228</u>	0.172	<u>0.380</u>	0.192	0.435
	CvT2DistilGPT2 [†] [15]	0.473	0.304	0.224	<u>0.175</u>	0.376	0.200	0.694
	R2GenGPT [†] [22]	<u>0.488</u>	0.316	<u>0.228</u>	0.173	0.377	<u>0.211</u>	0.438
	Ours (Modality-wise Fusion)	0.490	0.323	0.232	0.180	0.385	0.218	<u>0.491</u>

Comparison with the state-of-the-art Table 1 compares KARGEN’s performance with state-of-the-art (SOTA) methods in image captioning and report generation on the MIMIC-CXR and IU-Xray datasets. Table 2 focuses on comparisons using the metrics RadGraph F1, Bert score, and RadCliQ. Except those [†] marked methods whose performances are quoted from their respective papers, we re-run publicly released codes of comparison methods on the same training-test partition as our approach.

Table 2. Evaluation of Clinic-related Metrics on MIMIC-CXR

Methods	RadGraph F1(↑)	Bert Score(↑)	RadCliQ(↓)
R2Gen [3]	0.172	0.406	1.228
R2GenCMN [2]	0.182	0.418	1.182
CvT2DistilGPT2 [15]	0.196	0.374	1.220
RaDialog-RG [†] [17]	-	0.40	-
R2GenGPT [22]	0.187	0.415	1.207
Ours (Modality-wise Fusion)	0.203	0.421	1.165

As seen in Table 1, KARGEN outperforms existing methods across almost all evaluation metrics on both datasets. Specifically, it surpasses both traditional image captioning methods such as Show-Tell [25] and M2Transformer [4], advanced transformer-based R2Gen methods such as METransformer [21] and PPKED [11], and very recent LLM-based models like CvT2DistilGPT2 [15], RaDialog-RG [17], and R2GenGPT [22] in nearly all metrics. On MIMIC-CXR,

our BLEU-4 score sees a noteworthy improvement of 4.5%, rising from 0.134 to 0.140. Although our CIDEr score of 0.289 is lower than that of METransformer (0.362) and CvT2 (0.389), this discrepancy can be attributed to the employment of a unique expert voting in METransformer and the utilization of a larger image size (384x384 pixels) in CvT2. On IU-Xray, KARGEN consistently shows promising performance. In addition to NLG metrics, it is more important to see KARGEN achieve the highest scores in the clinic-related metrics RadGraph F1, Bert Score, and RadCliQ, reinforcing its advantages. This significant advancement is attributed to the integration of disease-related features, enhancing the model’s ability to accurately identify diseases. It is noted that RaDialog-RG [17] constructed prompts using the output of a trained disease classifier to incorporate disease information. Compared with it, our disease knowledge graph could carry more complicated disease relationships to assist LLMs for R2Gen.

Ablation Study: Table 3 summarizes our ablation study on the MIMIC-CXR dataset, singling out the contribution of each component, including knowledge-enhanced disease-related features, Graph Convolutional Network (GCN), and fusion methods. As seen, utilizing only regional or disease-related features yields moderate performance, while integrating both significantly enhances model effectiveness. Modality-wise fusion appears to be a superior fusion strategy. Examining configurations excluding GCN, which aggregates features through the graph, indicates less pronounced performance gains. Our complete model yields notably more accurate and descriptive outcomes compared to the baseline. Fig. 3 shows examples of generated reports. As seen, our model effectively captures both normal and abnormal contents consistent with the ground truth, while the baseline fails to generate the contents marked in red and magenta colors, confirming the benefits of our integration of knowledge-enhanced disease-related features via modality-wise fusion.

Table 3. Ablation study. $\mathbf{Z}_v$ is for regional features, and $\tilde{\mathbf{Z}}_g$ for disease-related features. **E**, **M**, and **A** stand for Element-wise, Modality-wise and Average.

Dataset	$\mathbf{Z}_v$	$\tilde{\mathbf{Z}}_g$	E-Fusion	M-Fusion	A-Fusion	GCN	BLEU-4	ROUGE	METEOR	CIDEr
MIMIC-CXR	✓						0.134	0.297	0.160	0.269
		✓					0.134	0.302	0.160	0.259
	✓	✓			✓	✓	0.132	0.303	0.156	0.245
	✓	✓	✓			✓	0.137	0.303	0.163	0.281
	✓	✓		✓			0.134	0.301	0.162	0.270
	✓	✓		✓		✓	0.140	0.305	0.165	0.289

4 Conclusions

In this paper, we propose a novel framework integrating LLMs with a medical knowledge graph for R2Gen. Our work highlights the value of incorporating

	Baseline	Ours (Modality-wise Fusion)	Ground Truth
	impression : no acute cardiopulmonary abnormality . findings : the patient is status post median sternotomy and cabg . heart size is normal . mediastinal and hilar contours are unremarkable . pulmonary vasculature is normal . lungs are clear . no focal consolidation pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities .	impression : no acute cardiopulmonary process . findings : left-sided port-a-cath tip terminates in the mid svc . lung volumes are low resulting in crowding of the bronchovascular structures . there is no focal consolidation pleural effusion or pneumothorax . the cardiomeastinal silhouette is within normal limits . median sternotomy wires and mediastinal clips are noted . no acute osseous abnormalities .	impression : no acute cardiopulmonary process . findings : there are low lung volumes . the lungs are clear . there is no pleural effusion or pneumothorax . the cardiomeastinal silhouette is unremarkable . left central line terminates in the right atrium . median sternotomy wires and mediastinal clips are noted . a calcified lymph node is noted in the ap window .
	impression : no acute cardiopulmonary process . findings : ap and lateral views of the chest . the lungs are clear of consolidation effusion or pulmonary vascular congestion . the cardiomeastinal silhouette is within normal limits . atherosclerotic calcifications noted at the aortic arch . no acute osseous abnormalities identified .	impression : no acute cardiopulmonary process . findings : frontal and lateral views of the chest were obtained . the lungs are hyperinflated with flattening of the diaphragms suggesting chronic obstructive pulmonary disease . there is no focal consolidation pleural effusion or pneumothorax . the cardiac silhouette is top-normal to mildly enlarged . atherosclerotic calcifications are noted at the aortic arch . no acute osseous abnormality	impression : stable mediastinal contour which is not widened . findings : frontal and lateral views of the chest were obtained . lungs are hyperinflated flattening of the diaphragms suggesting chronic obstructive pulmonary disease . 7-mm calcific focus in the left mid chest is stable . cardiac silhouette top normal to mildly enlarged . the aorta is tortuous .

Fig. 3. Examples of the generated reports. For better illustration, different colours highlight different medical terms in the reports.

disease-specific knowledge graphs with LLMs and the importance of fusing regional image features with knowledge-enhanced disease-related features to improve the quality and clinic utility of the generated reports. In the future, larger knowledge graphs will be explored along this line.

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

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. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
3. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
4. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
5. 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)
6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
7. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. 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)
9. 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)
 10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
 11. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
 12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
 13. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 375–383 (2017)
 14. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. *Artificial Intelligence Review* **42**, 275–293 (2014)
 15. Nicolson, A., Dowling, J., Koopman, B.: Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine* **144**, 102633 (2023)
 16. 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)
 17. Pellegrini, C., Özsoy, E., Busam, B., Navab, N., Keicher, M.: Radialog: A large vision-language model for radiology report generation and conversational assistance. arXiv preprint arXiv:2311.18681 (2023)
 18. 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)
 19. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
 20. Wang, Y., Lin, Z., Dong, H.: Rethinking medical report generation: Disease revealing enhancement with knowledge graph. arXiv preprint arXiv:2307.12526 (2023)
 21. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023)
 22. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology* **1**(3), 100033 (2023)
 23. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022)
 24. Wang, Z., Zhou, L., Wang, L., Li, X.: A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2433–2442 (2021)
 25. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual

- attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
26. 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)
 27. Yin, C., Qian, B., Wei, J., Li, X., Zhang, X., Li, Y., Zheng, Q.: Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In: 2019 IEEE international conference on data mining (ICDM). pp. 728–737. IEEE (2019)
 28. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. *Patterns* **4**(9) (2023)
 29. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019)
 30. 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)