TRRG: Towards Truthful Radiology Report Generation With Cross-modal Disease Clue Enhanced Large Language Models

Yuhao Wang[1], Yue Sun[2], Tao Tan[2], Chao Hao[1], Yawen Cui[1], Xinqi Su[1], Weichen Xie[3,5], Linlin Shen[3,4,5], and Zitong Yu[1,4,5,6⋆]

[1] School of Computing and Information Technology, Great Bay University, China
[2] Faculty of Applied Sciences, Macao Polytechnic University, China
[3] School of Computer Science and Software Engineering, Shenzhen University, China
[4] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, China
[5] Guangdong Provincial Key Laboratory of Intelligent Information Processing & Shenzhen Key Laboratory of Media Security, Shenzhen University, China
[6] Dongguan Key Laboratory for Intelligence and Information Technology

Abstract. The vision-language capabilities of multi-modal large language models have gained attention, but radiology report generation still faces challenges due to imbalanced data distribution and weak alignment between reports and radiographs. To address these issue, we propose TRRG, a stage-wise training framework for truthful radiology report generation. In the pre-training stage, contrastive learning enhances the visual encoder's ability to capture fine-grained disease details. In the fine-tuning stage, our clue injection module improves disease perception by integrating robust zero-shot disease recognition. Finally, the cross-modal clue interaction module enables effective multi-granular fusion of visual and disease clue embeddings, significantly improving report generation and clinical effectiveness. Experiments on IU-Xray and MIMIC-CXR show that TRRG achieves state-of-the-art performance, enhancing disease perception and clinical utility.

Keywords: Radiology Report Generation · Large Language Model · Chest X-ray

1 Introduction

The rapid development of machine learning, combined with the proliferation of large-scale public datasets, has driven significant progress in automated radiology report generation. This technology aims to assist radiologists by reducing workload, minimizing diagnostic errors, and streamlining clinical workflows. Radiology report generation involves producing detailed and structured textual descriptions for medical images such as X-rays, MRI, and CT scans. These reports must not only convey normal findings, but also capture subtle abnormal

⋆ Corresponding author: `zitong.yu@ieee.org` observations—such as disease categories, lesion locations, and severity indicators—that are essential for clinical decision making. Compared to general image captioning [24, 18, 21, 19, 10], radiology report generation presents unique challenges, including sparse and noisy disease supervision, a need for fine-grained abnormality localization, and the risk of generating clinically irrelevant or incomplete descriptions. Many recent encoder-decoder methods [12, 4, 30, 35, 28, 36, 11, 16] are limited by their reliance on coarse-grained visual-text alignment and struggle to achieve clinically meaningful disease perception, restricting their reliability in real-world applications.

Recent advances in multimodal learning and high-resolution feature representation further highlight opportunities and challenges in this field. For instance, FusionMamba [34] proposes dynamic feature enhancement for multimodal image fusion, offering insights into how multi-source feature interaction can enrich downstream tasks, including radiology reporting. Additionally, pre-training on high-resolution X-ray images, as studied in [27], demonstrates that rich visual detail is crucial for capturing subtle pathologies and improving diagnostic model performance. These developments suggest that more effective report generation frameworks should leverage advanced cross-modal interaction mechanisms and high-resolution representations to better capture disease-relevant clues. To address these challenges, we introduce **TRRG**, a novel framework for radiology report generation that integrates disease clue injection with powerful multi-modal learning. TRRG employs a stage-wise, cross-modal strategy, incorporating contrastive pretraining to refine the disease sensitivity of vision encoders. In the fine-tuning stage, we inject visual disease tokens and semantic clue embeddings to provide richer and more targeted supervision. A dedicated cross-modal clue interaction module further aligns visual and textual features, enabling the model to attend to clinically relevant image regions and facilitating more precise disease characterization. Inspired by the findings of [34], our approach dynamically enhances multi-source clues for robust report generation. Furthermore, motivated by [27], we utilize high-resolution representations to ensure subtle pathologies are effectively captured and described. We also propose a disease-aware consistency loss, which enforces alignment between vision and clue embeddings, reinforcing the model's ability to generate accurate, disease-focused reports.

We evaluate TRRG on the IU-Xray [6] and MIMIC-CXR [9] datasets, demonstrating that our approach surpasses existing methods in both language generation quality and clinical relevance. Extensive ablation studies validate the individual contributions of disease clue injection, cross-modal interaction, and consistency loss.

Our main contributions are summarized as follows:

- We propose TRRG, a disease clue injection enhanced large language model for radiology report generation, which addresses coarse-grained visual-text alignment and empowers the model to achieve fine-grained, disease-aware perception.
- We develop a cross-modal disease clue interaction module that effectively integrates visual embeddings and disease clue embeddings, guiding large

language models to generate higher-quality, clinically meaningful radiology reports.

- Drawing on recent findings in multimodal fusion [34] and high-resolution pre-training [27], our framework leverages dynamic feature enhancement and detailed visual representations to further boost clinical performance.
- Comprehensive experiments on IU-Xray and MIMIC-CXR confirm that TRRG achieves state-of-the-art results in both language generation quality and clinical reliability, with ablation studies clarifying the impact of each core module.

2 Methods

The proposed TRRG follows a two-stage training process. In the pre-training stage, we enhance disease-aware fine-grained alignment between radiographs and reports using sentence-level contrastive learning, improving the vision encoder's disease perception. In the fine-tuning stage, the clue injection and cross-modal clue interaction modules further refine language generation and clinical effectiveness by aligning visual and disease clue embeddings. These components significantly boost both alignment and report quality. The details of TRRG are shown in Fig. 1.

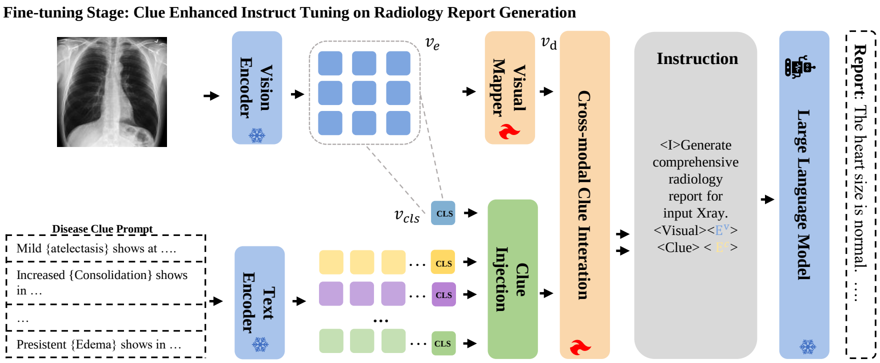

Fine-tuning Stage: Clue Enhanced Instruct Tuning on Radiology Report Generation
Vision Encoder
v_e
Visual Mapper
v_d
Cross-modal Clue Interation
Instruction
<I>Generate comprehensive radiology report for input Xray.
<Visual><E^v>
<Clue> < E^c>
Large Language Model
Report: The heart size is normal. ...
v_{cls}
CLS
Disease Clue Prompt
Mild {atelectasis} shows at ....
Increased {Consolidation} shows in ...
...
Presistent {Edema} shows in ...
Text Encoder
Clue Injection

Fig. 2. During fine-tuning, the frozen visual and clue encoders inject disease clues via the clue injection module, enabling cross-modal interaction, while a frozen large language model undergoes instruction-based fine-tuning for medical report generation.

2.1 Stage 1: Disease-aware Cross-modal Fine Grained Alignment

Recent studies [33, 25] have shown that pre-trained CLIP models trained on large-scale

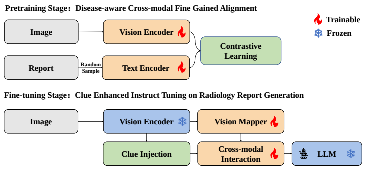


Fig. 1. The training strategy of our proposed TRRG

medical image-text pairs can achieve human-level accuracy in zero-shot disease classification tasks. Inspired by this, we adopt a stage-wise training approach. In the pre-training stage, we randomly sample sentences from radiology reports to train the model, thereby enhancing the vision encoder's representation capability and supporting robust disease clue prompting.

Given an image-text pair (I, T), where $T = \{T_1, \ldots, T_t\}$ and each T_i is a sentence from the report, the image is encoded as $\mathbf{v} = \{\mathbf{v}_{cls}, \mathbf{v}_1, \ldots, \mathbf{v}_n\} \in \mathbb{R}^{(n+1)\times d}$ by a transformer-based vision encoder, and a randomly sampled sentence T_r is encoded as $\mathbf{t} = \{\mathbf{t}_{cls}, \mathbf{t}_1, \ldots, \mathbf{t}_m\} \in \mathbb{R}^{(m+1)\times d}$ by a BERT-based model, where "cls" denotes the pooling token. We use the "CLS" tokens, i.e., $\mathbf{v}_{cls}$ and $\mathbf{t}_{cls}$, as the global representations, i.e., $\mathbf{v} = \mathbf{E}_{img}(I)$ and $\mathbf{t} = \mathbf{E}_{txt}(T_r)$, with T_r being a randomly selected sentence.

Contrastive loss is then computed to align pooled image and text embeddings. Specifically, for paired image and text embeddings $\{\mathbf{v}'_i, \mathbf{t}_i\}_{i=1}^N$, where N is the batch size and τ is a temperature parameter, the loss is:

$$\mathcal{L} = -\left(\log \frac{\exp(\sigma(\mathbf{t}_i, \mathbf{v}'_i)/\tau)}{\sum_{j=1}^N \exp(\sigma(\mathbf{t}_i, \mathbf{v}'_j)/\tau)} + \log \frac{\exp(\sigma(\mathbf{v}'_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^N \exp(\sigma(\mathbf{v}'_i, \mathbf{t}_j)/\tau)}\right) \tag{1}$$

This contrastive learning stage improves the disease-oriented visual representation, benefiting subsequent report generation.

2.2 Stage 2: Clue Enhanced Instruct Tuning on Radiology Report Generation

In this stage, we enhance report generation by integrating disease clue prompts. The image is first encoded as $\mathbf{v}_e = E_{\text{img}}(x)$, then mapped to the visual space via a trainable mapper: $\mathbf{v}_d = W\mathbf{v}_e + b$, where W is the trainable weight. The disease visual expert token is computed by averaging patch tokens, i.e., $\mathbf{v}_{cls} = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i$, with each $\mathbf{v}_i \in \mathbb{R}^{1\times d}$. This yields both the disease visual embedding $\mathbf{v}_d \in \mathbb{R}^{n\times d}$ and the expert token $\mathbf{v}_{cls} \in \mathbb{R}^{1\times d}$.

For the clue injection module, following [7], we construct disease clue prompts using templates such as: `Clue: <severity> <disease> at <location>`. We manually define templates for m common diseases (e.g., opacity, pneumonia) and encode each clue prompt with a frozen text encoder into $\mathbf{c}^i = \{\mathbf{c}_{cls}, \mathbf{c}_1, ..., \mathbf{c}_r\} \in \mathbb{R}^{(r+1)\times d}$, where r is the prompt length and $i = 1, ..., m$.

To measure the relevance between visual and clue tokens, we compute clue weights by $w_i = \text{softmax}(\mathbf{v}_{cls} \cdot \mathbf{c}^i_{cls})$ for $i = 1, ..., m$, where $w \in \mathbb{R}^{1\times m}$ and $\mathbf{c}^i_{cls}$ is the expert token of the i-th clue. We then select the top-k clues with the highest weights as final inputs: $\mathbf{c}_s = \{\mathbf{c}^i \mid i \in \text{topk}(w, k)\}$, where $\text{topk}(w, k)$ returns the indices of the k largest weights. The resulting set $\mathbf{c}_s \in \mathbb{R}^{k\times r\times d}$ serves as the disease expert clues injected into the model for generation.

Cross Modal Clue Interaction For multi-disease clues $\mathbf{c}_s \in \mathbb{R}^{k\times r\times d}$, these clues align with frozen vision encoder representations but lack interaction with visual mapper features. To address this, we propose a Cross-Modal Clue Interaction Module to enhance generative representations and facilitate cross-modal interaction. Given the typically larger number of disease clue tokens, we introduce learnable queries to manage input redundancy. Additionally, we propose a Disease Clue Consistency Loss to maintain attention on disease clue embeddings and provide supervision for fine-tuning large language models. The multimodal module employs a two-stream architecture with self-attention for intra-modal interaction and cross-attention for aligning textual clues with visual representations, ensuring cross-modal consistency. For visual embedding, $\mathbf{v}_d = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$, disease clue embedding $\mathbf{c}_s \in \mathbb{R}^{k\times r\times d}$, We flatten the disease clues into clue tokens $\mathbf{c}_s = \{\mathbf{c}_1, ..., \mathbf{c}_{(k\times r)}\}$, where $\mathbf{c}_i \in \mathbb{R}^{1\times d}$. Next, both visual embeddings and clue embeddings are fed into linear projection layer and atttention layer:

$$\mathbf{V}' = \mathrm{Attn}(Q^v, K^v, V^v), \mathbf{C}' = \mathrm{Attn}(Q^c, K^c, V^c), \tag{2}$$

where $\mathbf{V}', \mathbf{C}'$ are visual embeddings and disease clue embeddings after multi-head self attention layer. Next, we utilize learnable tokens $E = \{E_1, E_2, ..., E_L\}$ as a common feature space to establish associations between the visual and textual modalities, where L represents the number of learnable tokens. In detail, we employ a scaled dot-product attention layer to calculate the correlation between the learnable tokens E and the mapped visual tokens E^v. We perform the same operation on learnable queries and disease clue embeddings and obtained clue tokens E^c. This process can be expressed as:

$$\mathrm{E}^v = \mathrm{FFN}(\mathrm{Attn}(E^e, \mathbf{V}', \mathbf{V}')), \mathrm{E}^c = \mathrm{FFN}(\mathrm{Attn}(E^e, \mathbf{C}', \mathbf{C}')). \tag{3}$$

Furthermore, since disease clues are often sparse during cross-modal interaction, to enhance the consistency between visual tokens and clue tokens and improve effective supervision signals during radiology report generation, we propose a disease-aware consistency loss. Our disease-aware consistency loss is calculated as:

$$\mathcal{L}_{DC} = -\frac{1}{K}\sum_{i=1}^{K}\frac{E^v \cdot E^c}{\|E^v\|\|E^c\|}, \tag{4}$$

We calculate the similarity between visual tokens and clue tokens and aim to maximize the alignment between visual and textual tokens. The disease-aware consistency loss effectively endows visual tokens with the ability to perceive diseases.

Optimization Objective Our overall objective function is:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{DC}. \tag{5}$$

where $\mathcal{L}_{CE}$ is the Cross Entropy loss which is most used in captioning task.

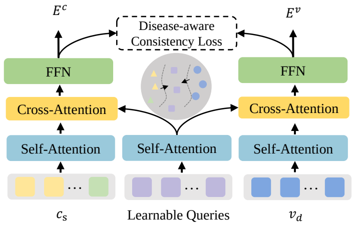


Fig. 3. Architecture of Cross Modal Clue Interaction Module

Table 1. Comparisons of the proposed TRRG with previous studies on the IU X-Ray and MIMIC-CXR test set with respect to language generation (NLG) and clinical efficacy (CE) metrics. Best results are in **bold**.

Dataset	Model	NLG Metrics							CE Metrics		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	Precision	Recall	F1
IU X-Ray	HGRG-Agent [14]	0.438	0.298	0.208	0.151	0.322	-	0.343	-	-	-
	KERP [11]	0.482	0.325	0.226	0.162	0.339	-	0.280	-	-	-
	R2Gen [4]	0.470	0.304	0.219	0.165	0.371	0.187	-	-	-	-
	PPKED [16]	0.483	0.315	0.224	0.168	0.376	0.187	0.351	-	-	-
	GSK [35]	**0.496**	**0.327**	**0.238**	**0.178**	**0.381**	-	0.382	-	-	-
	R2GenCMN [3]	0.475	0.309	0.222	0.170	0.375	0.191	-	-	-	-
	METransformer [31]	0.483	0.322	0.228	0.172	0.380	0.192	**0.435**	-	-	-
	TRGG (Ours)	0.482	0.302	0.217	0.151	0.377	**0.209**	0.405	-	-	-
MIMIC-CXR	M2Transformer [5]	0.332	0.210	0.142	0.101	0.264	0.134	0.142	-	-	-
	R2Gen [4]	0.353	0.218	0.145	0.103	0.277	0.142	-	0.333	0.273	0.276
	PPKED [16]	0.36	0.224	0.149	0.106	0.284	0.149	0.237	-	-	-
	GSK [35]	0.363	0.228	0.156	0.115	0.284	-	0.203	-	-	-
	R2GenCMN [3]	0.353	0.218	0.148	0.106	0.278	0.142	-	0.334	0.275	0.278
	MSAT [29]	0.373	0.235	0.162	0.120	0.282	0.143	0.299	-	-	-
	METransformer [31]	0.386	0.250	0.169	0.124	0.291	0.152	**0.362**	0.364	0.309	0.311
	DCL [13]	-	-	-	0.107	0.284	0.150	0.281	**0.471**	0.352	0.373
	R2GenGPT [32]	0.365	0.237	0.163	0.117	0.277	0.136	0.145	0.341	0.312	0.325
	FGIRG [2]	0.379	0.234	0.154	0.106	0.285	0.162	-	-	-	-
	R2GMMN [22]	0.396	0.244	0.162	0.115	0.274	0.151	-	0.411	0.398	0.389
	TRGG (Ours)	**0.436**	**0.298**	**0.213**	**0.157**	**0.336**	**0.167**	0.219	0.403	**0.399**	**0.393**

3 Experiment

3.1 Datasets and Evaluation Metrics

IU-Xray [6] is a widely recognized benchmark dataset for radiology report generation. The dataset consists of over 7,470 chest X-ray images and 3,955 corresponding radiology reports manually annotated by expert radiologists. **MIMIC-CXR** [9] is a dataset comprising 64,588 patients collected at the Beth Israel Deaconess Radiology Center between 2011 and 2016. It includes 77,110 chest X-ray images and 227,835 corresponding free-text radiology reports. To ensure experimental fairness, we replicated the experimental settings of previous studies. This led to a training set of 222,758 samples, with validation and test sets comprising 1,808 and 3,269 samples, respectively. Based on previous research, we evaluate our proposed radiology report generation model from two perspectives. 1) Evaluation of Language Generation Quality (**NLG Metrics**): Utilizing commonly used linguistic evaluation metrics such as BLEU [20], Rouge-L [15], and CIDEr [26]. 2) Clinical Effectiveness Metrics (**CE Metrics**): We employ NLP text disease labeler ChexBERT[23] for text classification. We extract 14 common diseases from the generated reports and reference report. Precision, recall, and F1 score are used to assess performance in terms of clinical efficacy.

4 Main Results

Implementation Details We utilized Mistral-7B [8] as the language model, Swin-Transformer [17] as the vision encoder, and ClinicalBERT [1] for text encoding during pretraining. Fourteen common diseases were selected for template construction, with a maximum of three injected disease clues (K=3), and both visual and text embeddings set to 1024 dimensions. The cross-modal cue interaction module adopted an 8-head attention mechanism, each with a single attention and cross-attention layer. Training was conducted on four NVIDIA A40 48GB GPUs using the MIMIC-CXR dataset (5 epochs) and IU-Xray (20 epochs), with a batch size of 8 and a learning rate of 1e-4.

We benchmarked our model against state-of-the-art image captioning and radiology report generation methods, including R2Gen [4], R2GenCMN [3], PPKED [16], R2GenGPT [32], FGIRG [2], and R2GMMN [22]. As shown in Table 2.2, our model achieves superior NLG performance, particularly on MIMIC-CXR, benefiting from the larger dataset for enhanced pretraining and cross-modal alignment. Although METransformer [31] achieves higher CIDER scores due to its optimization strategy, our model excels in language consistency, semantic richness, and accurate radiology report generation. In clinical efficacy evaluation, our model outperforms fine-tuning methods such as R2GenGPT, achieving average accuracy, recall, and F1 scores of 0.403, 0.399, and 0.393, respectively, surpassing R2GMMN and demonstrating improved disease recognition and report accuracy.

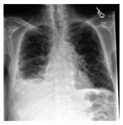
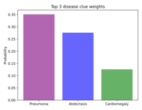
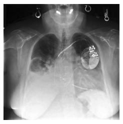
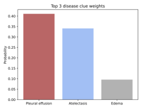
	**Ground-Truth:** There is no clear radiographic change over the past 11 days. **Bilateral pleural effusions moderate** on the right small on the left and callus **pulmonary nodules** are unchanged. **Confluent opacification** at the base of the right lung **is probably atelectasis**, **pleural mild pneumonia** is difficult to exclude.	**Baseline:** Frontal and lateral views of the chest. The pulmonary vasculature is normal. The lungs are clear without consolidation, effusion or pneumothorax. The cardiomediastinal silhouette is normal.	**Ours**: The cardiomediastinal silhouette and pulmonary vasculature are within normal limits . **Pneumonia at the right lung base is not excluded.** there is no focal airspace consolidation pleural effusion or pneumothorax . there is no acute bony abnormality.	Top 3 Clue Weights: 
	**Ground-Truth:** Since the prior study, there is no change in **large right pleural effusion** and **associated atelectasis**. Heart size and mediastinum are unchanged including cardiomegaly. **Biventricular pacer is redemonstrated**	**Baseline:** Frontal views of the chest. Left chest wall **pacing device** seen with leads in the right atrium. The lungs are clear, and the cardiomediastinal silhouette and hila are normal. **pleural effusions show in in right lung zone.**	**Ours**: the heart size is normal. the mediastinal and hilar contours are normal. **Low lung volumes with probable bibasilar atelectasis**. There is no pneumothorax. Pulmonary vascularity is normal. **there are small bilateral pleural effusions.** Cardiomegaly is unchanged	Top 3 Clue Weights: 

Fig. 4. We compare the generated results of the base model and the TRRG (Ours) with the ground truth, highlighting key information using colored fonts. Our model effectively generated specific descriptions tailored to diseases.

4.1 Ablation Study

Effectiveness of each component. We constructed our baseline model, "BASE," by fine-tuning only the visual mapper. We then introduced a disease clue injection module (DCI), a cross-modal clue interaction module (CMCI), and a disease-aware consistency loss function (DAL). The " + " symbol in Table 2 represents the effects of adding these components. Each module significantly enhances performance, with systematic partitioning ensuring comparability despite some experimental randomness. Our proposed modules achieve 14.5%, 15.8%, 13.6%, 11.7%, and 9.9% improvements in BLEU-4, ROUGE-L, METEOR, CIDEr, and F1 scores from "BASE" to TRRG, validating our approach.

Table 2. Ablation study of different componet we proposed, "DCI," "CMCI," and "DAL" respectively denote the Disease Clue Injection module, Cross-Modal Clue Interaction module, and Disease-Aware Loss function.

Models	IU-Xray				
	BLEU-4	ROUGE	METEOR	CIDER	F1
BASE	0.156	0.370	0.194	0.387	-
BASE+DCI	0.152	0.365	0.197	0.390	-
BASE+DCI+CMCI	0.147	0.372	0.205	0.402	-
TRRG	0.151	0.377	0.209	0.405	-
	MIMIC-CXR				
BASE	0.137	0.290	0.147	0.196	0.354
BASE+DCI	0.142	0.311	0.156	0.207	0.384
BASE+DCI+CMCI	0.159	0.324	0.162	0.211	0.387
TRRG	0.157	0.336	0.167	0.219	0.393

4.2 Qualitative analysis

We conduct a qualitative analysis to validate the effectiveness of the proposed model. As depicted in Fig 4, our disease clues and probabilities are highlighted using colored fonts. Compared to the base model, our proposed model tends to include more disease-related content with injected clues during the process of generating radiology reports. This observation confirms that our model has higher clinical reliability. In the second example, the conventional model hinted at the presence of auxiliary devices in the image, but our model failed to provide relevant descriptions. This indicates some limitations of our proposed approach, which may lead to an overemphasis on disease-related content in certain cases, thereby compromising findings and obscuring the expression of basic descriptions.

5 Conclusion

In this paper, we propose the TRGG for truthful radiology report generation based on fine-tuning large language models with injected disease cues. Our proposed stage-wise training strategy effectively promotes cross-modal alignment between radiography and reports. The clue injection module and cross-modal clue interaction module proposed by us can effectively facilitate the semantic representation of diseases and cross-modal alignment. Experimental results demonstrate the superiority of our approach. Future research directions include developing a generalizable method for medical image report generation that can be applied across various medical imaging text report datasets, enabling further extension to heterogeneous modalities such as CT, MRI, and Ultrasound.

6 Disclosure of Interests

The authors declare no competing interests.

7 Acknowledgement

The work was supported by the National Natural Science Foundation of China under grant no. 82261138629, 12326610, and 62276170. This work is also supported by Science and Technology Development Fund of Macao (0041/2023/RIB2), Open Fund of National Engineering Laboratory for Big Data System Computing Technology (Grant No. SZU-BDSC-OF2024-02), Guangdong Provincial Key Laboratory (Grant 2023B1212060076), Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037). This work was also supported by Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

[1] Emily Alsentzer et al. “Publicly available clinical BERT embeddings”. In: *arXiv preprint arXiv:1904.03323* (2019).

[2] Wenting Chen et al. “Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation”. In: *62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. 2024.

[3] Zhihong Chen et al. “Cross-modal memory networks for radiology report generation”. In: 2022.

[4] Zhihong Chen et al. “Generating radiology reports via memory-driven transformer”. In: *EMNLP*. 2020.

[5] Marcella Cornia et al. “Meshed-Memory Transformer for Image Captioning”. In: *CVPR*. 2020.

[6] Demner Fushman Dina et al. “Preparing a collection of radiology examinations for distribution and retrieval”. In: *Journal of the American Medical Informatics Association Jamia* (2015).

[7] Shih-Cheng Huang et al. “GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition”. en-US. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021. DOI: `10.1109/iccv48922.2021.00391`. URL: `http://dx.doi.org/10.1109/iccv48922.2021.00391`.

[8] Albert Q Jiang et al. “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).

[9] Alistair E. W Johnson et al. “MIMIC-CXR: A large publicly available database of labeled chest radiographs”. In: *CoRR* (2019).

[10] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. “DenseCap: Fully Convolutional Localization Networks for Dense Captioning”. In: *CVPR*. 2016.

[11] Christy Y. Li et al. “Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation”. In: *AAAI*. 2019.

[12] Christy Y. Li et al. *Knowledge-driven Encode, Retrieve, Paraphrase for Medical Image Report Generation.* 2019.

[13] Mingjie Li et al. “Dynamic Graph Enhanced Contrastive Learning for Chest X-ray Report Generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3334–3343.

[14] Yuan Li et al. “Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation”. In: *NeurIPS*. 2018.

[15] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *ACL*. 2004.

[16] Fenglin Liu et al. “Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation”. In: *CVPR*. 2021.

[17] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[18] Jiasen Lu et al. “Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning”. In: *CVPR*. 2017.

[19] Yingwei Pan et al. “X-Linear Attention Networks for Image Captioning”. In: *CVPR*. 2020.

[20] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[21] Steven J. Rennie et al. “Self-Critical Sequence Training for Image Captioning”. In: *CVPR*. 2017.

[22] Hongyu Shen et al. “Automatic Radiology Reports Generation via Memory Alignment Network”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 5. 2024, pp. 4776–4783.

[23] Akshay Smit et al. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. 2020. arXiv: `2004.09167 [cs.CL]`.

[24] Mingkang Tang et al. “Clip4caption: Clip for video caption”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 4858–4862.

[25] Ekin Tiu et al. "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning". In: *Nature Biomedical Engineering* 6.12 (2022), pp. 1399–1406.

[26] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "CIDEr: Consensus-based image description evaluation". In: *CVPR*. 2015.

[27] Xiao Wang et al. "Pre-training on high-resolution X-ray images: an experimental study". In: *Visual Intelligence* 3.1 (2025), pp. 1–15.

[28] Yuhao Wang et al. "Self adaptive global-local feature enhancement for radiology report generation". In: *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2023, pp. 2275–2279.

[29] Zhanyu Wang et al. "A medical semantic-assisted transformer for radiographic report generation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 655–664.

[30] Zhanyu Wang et al. "A Self-Boosting Framework for Automated Radiographic Report Generation". In: *CVPR*. 2021.

[31] Zhanyu Wang et al. "METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens". In: *CVPR*. 2023, pp. 11558–11567.

[32] Zhanyu Wang et al. "R2gengpt: Radiology report generation with frozen llms". In: *Meta-Radiology* 1.3 (2023), p. 100033.

[33] Zifeng Wang et al. "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text". en-US. In: (Oct. 2022).

[34] Xinyu Xie et al. "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba". In: *Visual Intelligence* 2.1 (2024), p. 37.

[35] S. Yang et al. "Knowledge Matters: Radiology Report Generation with General and Specific Knowledge". In: *Medical Image Analysis* (2021).

[36] Changchang Yin et al. "Automatic Generation of Medical Imaging Diagnostic Report with Hierarchical Recurrent Neural Network". In: *ICDM*. 2020.