Contrastive Knowledge-Guided Large Language Models for Medical Report Generation

Yuyang Sha, Hongxin Pan, Weiyu Meng, Kefeng Li✉

Faculty of Applied Sciences, Macao Polytechnic University
kefengl@mpu.edu.mo

Abstract. Automatic medical report generation (MRG) holds considerable research value and has the potential to significantly alleviate the workload of radiologists. Recently, the rapid development of large language models (LLMs) has improved the performance of MRG. However, numerous challenges still need to be addressed to achieve highly accurate medical reports. For instance, most existing methods struggle to interpret image details, lack relevant medical knowledge, and overlook fine-grained cross-modality alignment. To overcome these limitations, we propose a knowledge-guided vision-language alignment framework with contrastive learning and LLMs for medical report generation. The designed method leverages visual representations, relevant medical knowledge, and enhanced features to generate accurate reports via the LLMs-based decoder. To improve the integration of medical-related information, we introduce the Knowledge Injection Module, which enhances the model's feature representation capabilities while unlocking medical domain knowledge in LLMs. Inspired by the contrastive learning scheme, we introduce the Contrastive Alignment Module to align the visual features and textual information effectively. Additionally, the Cross-Modality Enhancement Module can retrieve similar reports for the input images to boost diagnostic accuracy. We conduct extensive experiments on two popular benchmark datasets, including IU X-Ray and MIMIC-CXR. The results demonstrate that our proposed method achieves promising performance compared with state-of-the-art frameworks.

Keywords: Medical Report Generation · Large Language Models · Cross-Modality Alignment · Knowledge Graph · Contrastive Learning.

1 Introduction

Medical imaging plays a crucial role in modern healthcare, profoundly influencing the diagnosis and treatment of various diseases [22]. With the development of artificial intelligence technologies, automated analysis of medical images has gained widespread attention and achieved remarkable progress [18, 9, 20]. Medical report generation (MRG) is a crucial task in the automated analysis of medical images, which aims to deliver a coherent and accurate summary of the visual information present in these images. Compared to other tasks, MRG presents notable complexities and challenges. Given its potential to reduce the substantial

workload of radiologists, numerous approaches [12, 1, 10, 16] have been proposed to enhance the MRG performance.

Generating fluent and accurate medical reports automatically remains a challenging task. Algorithms need to comprehend both the overarching and detailed aspects of medical images while also possessing relevant medical knowledge. To address these problems, researchers have proposed various solutions. For instance, some studies [15, 14] try to utilize the knowledge graph to embed prior medical knowledge into the frameworks for report generation. Other researches [12, 1, 3] have made efforts to tackle these issues by incorporating disease classification subtasks, designing memory-driven modules, or implementing innovative report-generation manner. Recently, large language models (LLMs) have exhibited human-like cognitive abilities and skills in text generation [23, 17], revolutionizing various fields, such as chatbots [19] and medical diagnostics [21]. Consequently, many researchers try to apply LLMs to the MRG task. Li *et al.* [15] introduced KARGEN, which combines the LLMs with a disease knowledge graph to generate medical reports. Similarly, Wang *et al.* [25] proposed R2GenGPT, which leverages a visual-language model to efficiently generate diagnosis reports. While these LLMs-based methods have shown promising results, they still fall short in addressing the aforementioned challenges. Therefore, current MRG solutions may be unreliable for diagnosing rare diseases, severely reducing their clinical value.

In this paper we propose KACL, a **K**nowledge-guided vision-language **A**lignment framework with **C**ontrastive learning and **L**LMs to generate precise medical reports. The proposed KACL contains five components: the Visual Encoder, LLMs-based Decoder, Knowledge Injection Module (KIM), Contrastive Alignment Module (CAM), and Cross-Modality Enhancement Module (CEM). Among them, the Visual Encoder is responsible for extracting visual features from medical images. By utilizing visual representations and medical-related knowledge, the LLMs-based encoder generates corresponding diagnostic reports for the input samples. The KIM can inject medical-related knowledge into the proposed framework, enhancing the model's ability to derive powerful feature representations while unlocking relevant medical domain insights within the LLMs. Inspired by the contrastive learning scheme, we introduce CAM to align the visual and textual features, ultimately improving the accuracy of the generated reports. To further enhance diagnostic precision, we develop the CEM, which assists in diagnosing the query image by utilizing a pre-trained medical CLIP [7] model to retrieve similar reports from an external database. We summarize the key contributions as follows.

1. We propose a novel MRG solution based on LLMs named KACL. This approach leverages enhanced visual features, medical knowledge, and retrieved textual representations to form prompt tokens, guiding the LLMs-based decoder to produce precise and coherent medical reports.
2. The proposed KACL contains three key modules: Knowledge Injection Module (KIM), Contrastive Alignment Module (CAM), and Cross-Modality Enhancement Module (CEM). The KIM integrates medical-related knowledge

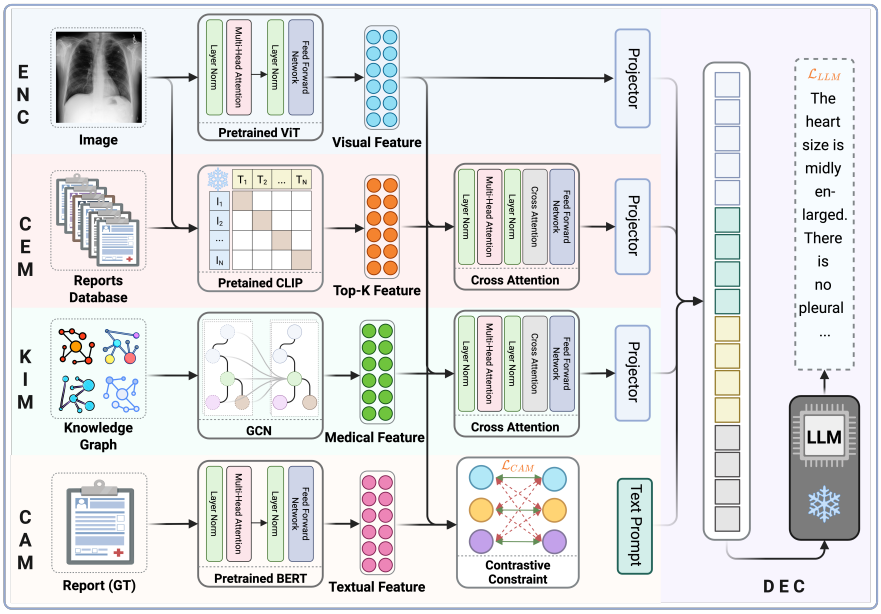

Fig. 1. The overall architecture of our proposed method. This framework mainly consists five components: Visual Encoder (ENC), LLMs-based Decoder (DEC), Knowledge Injection Module (KIM), Contrastive Alignment Module (CAM), and Cross-Modality Enhancement Module (CEM).

into the MRG framework. The CAM aligns the visual and textual features, while the CEM enhances diagnostic accuracy by retrieving similar reports as auxiliary information for medical images.

3. We conduct extensive experiments on two benchmark datasets: IU X-Ray and MIMIC-CXR. The results demonstrate that our method outperforms current state-of-the-art (SOTA) frameworks in most natural language generation (NLG) and clinical efficacy (CE) metrics.

2 Method

2.1 Framework

The proposed KACL, illustrated in Figure 1, comprises five components: Visual Encoder, LLMs-based Decoder, Knowledge Injection Module (KIM), Contrastive Alignment Module (CAM), and Cross-Modality Enhancement Module (CEM). Specifically, the input samples comprise a 2D medical image I, medical-related knowledge K, and external report databases R_E. The KACL aims to generate a report R_p for the input sample. The process can be formulated as: $R_p \leftarrow$

KACL(I, K, R_E). In this study, the pre-trained Vision Transformer (ViT) [6] is used as the Visual Encoder f_{ve} to extract latent representations X_i from the provided i_{th} medical image I_i, which can be defined as: $X_i = f_{ve}(I_i)$.

2.2 Knowledge Injection Module

To produce precise medical reports, the algorithm needs to carefully analyze image details and demonstrate a strong understanding of medical knowledge. Integrating medical-related knowledge into the proposed framework can enhance feature representation while unlocking insights of LLMs in the medical domain. Therefore, we design the Knowledge Injection Module (KIM) to inject real-world medical information into KACL. Inspired by CheXpert [11], we build a medical knowledge graph focusing on chest diseases, and its interconnection relationships are shown in Figure 2. Specifically, the proposed KIM comprises one cross-attention module and three graph convolutional network (GCN) layers. Initially, the extracted visual features X and encoded disease names E should be processed through the cross-attention module to obtain the initial node representations, which are formulated as follows.

$$N^0 = \text{MultiH-Att}(X, E) = \text{Concat}(h_1, h_2, ..., h_n)W^O, \tag{1}$$

where the MultiH-Att defines multi-head attention module, n is the number of heads. The h_i denotes the i_{th} heads, formulated as: $h_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i$. The $Q_i = EW_i^Q$, $K_i = XW_i^K$, and $V_i = XW_i^V$ represent the transformed query, key, and value. And the W_i^Q, W_i^K, W_i^V, and W^O are learnable parameter matrices.

Next, we leverage the designed KIM module to update the obtained features N^0. In summary, the representation rule of the l_{th} layer is:

$$N^{l+1} = \theta(A^l N^l W^l), \tag{2}$$

where the A^l defines the adjacency matrix, N^l refers the node features in the l_{th} GCN layer, and W^l indicates the trainable parameters. The $\theta(\cdot)$ serves as an activation function. Finally, we obtain the medical-related knowledge features $K = N^3$.

2.3 Contrastive Alignment Module

To investigate the relationship between extracted visual features and textual representation of medical reports, we develop the Contrastive Alignment Module (CAM) based on the contrastive learning scheme. It facilitates cross-modality semantic alignment between visual and textual information, enhancing the model's ability to obtain more discriminative representations. We utilize the f_{ve} to process the input image, selecting the Classify (CLS) Token as its overall representation. Subsequently, we employ a pre-trained medical BERT [5] model to convert the textual information from the ground truth medical report into another CLS Token. Then, the image and text representations are projected into a

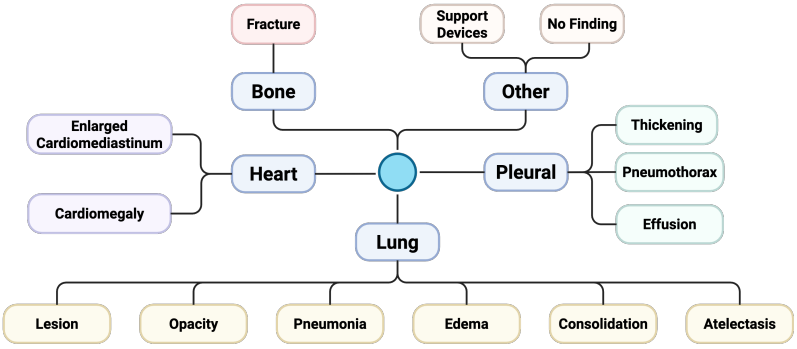

Fig. 2. The pre-constructed medical knowledge graph, where connected diseases are interrelated.

unified dimensional space utilizing linear projection. Finally, we can derive the image-to-text contrastive loss $\mathcal{L}_{i2t} = -\log \frac{\exp(v_i,t_i)/\tau}{\sum_{k=i}^{K}\exp(v_i,t_k)/\tau}$ and text-to-image contrastive loss $\mathcal{L}_{t2i} = -\log \frac{\exp(t_i,v_i)/\tau}{\sum_{k=i}^{K}\exp(t_i,v_k)/\tau}$. The v_i and t_i denote the extracted visual and textual representations, respectively. The τ represents a temperature hyperparameter, and K refers the batch size. The contrastive loss for CAM is:

$$\mathcal{L}_{CAM} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}). \tag{3}$$

The proposed CAM can effectively reduce the distance between positive pairs while increasing the distance between negative pairs. It is beneficial for generating more accurate reports.

2.4 Cross-Modality Enhance Module

Radiologists usually refer to some relevant documents when creating new diagnosis reports. Therefore, obtaining knowledge from relevant external databases can significantly boost the model's effectiveness and diagnostic accuracy in generating reports. In light of this concept, we propose the Cross-Modality Enhance Module (CEM), which employs the pre-trained CLIP [7] model to facilitate cross-modal retrieval for the provided medical images. Specifically, for each sample I, we identify the *Top-K* most relevant retrieved textual representations $R_E = \{r_E^1, r_E^2, ..., r_E^k\}$ from the external database. These retrieved representations R_E are processed through self-attention and subsequently cross-attended with visual feature v^i. Notably, the v^i serves as the query, while the output of the self-attention module acts as both the key and value. The enhanced cross-modality feature X^C is defined as:

$$X^C = \text{Cross-Att}(\text{Self-Att}(R_E), v^i). \tag{4}$$

Remarkably, the self-attention (Self-Att) and cross-attention (Cross-Att) modules are trainable, while the CLIP model remains frozen during training.

Table 1. Comparison results with state-of-the-art methods on IU X-Ray and MIMIC-CXR datasets in terms of NLG and CE metrics. The highest and the second-highest results are denoted in **bold** and underlines. RG-L, PREC., REC., and F1. represent ROUGE-L, Precision, Recall, and F1 Score, respectively.

Dataset	Method	NLG Metrics						CE Metrics		
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	RG-L	METEOR	PREC.	REC.	F1.
IU X-Ray	R2Gen [2]	0.470	0.304	0.219	0.165	0.371	0.187	-	-	-
	METrans [24]	0.483	0.322	0.228	0.172	0.380	0.192	-	-	-
	R2GenGPT [25]	0.488	0.316	0.228	0.173	0.377	0.211	-	-	-
	PromptMRG [12]	0.401	-	-	0.098	0.281	0.160	-	-	-
	BoostRRG [16]	0.499	0.323	0.238	**0.184**	**0.390**	0.208	-	-	-
	Ours	**0.501**	**0.326**	**0.244**	**0.184**	0.385	**0.211**	-	-	-
MIMIC CXR	R2Gen [2]	0.353	0.218	0.145	0.103	0.277	0.142	0.333	0.273	0.276
	METrans [24]	0.386	0.250	0.169	0.124	0.291	0.152	0.364	0.309	0.334
	R2GenGPT [25]	0.411	0.267	**0.186**	0.134	0.297	0.160	0.392	0.387	0.389
	PromptMRG [12]	0.398	-	-	0.112	0.291	**0.175**	0.501	**0.509**	**0.476**
	BoostRRG [16]	0.402	0.262	0.180	0.128	0.291	**0.175**	0.465	0.482	0.473
	Ours	**0.414**	**0.270**	0.184	**0.136**	**0.303**	0.169	**0.503**	0.442	0.469

2.5 Report Generation

The outputs produced by the Visual Encoder, KIM, and CEM are combined to form the fused feature X. Then, X and textual prompt P are fed into the LLMs-based decoder to facilitate the medical report generation. We employ the LLaMA3.1-8B [8] as the decoder. The decoding process can be outlined as follows:

$$r_t^p = f_{dec}(X, P, r_{1:t-1}^p). \tag{5}$$

where $r_t^p \in \mathbb{V}$ is the predict token at the step t, and $\mathbb{V}$ defines the vocabulary. The predicted report can be expressed as $R^p = \{r_1^p, r_2^p, ..., r_T^p\}$, and T is the length of the report. The language modeling loss is defined as follows:

$$\mathcal{L}_{LLM} = -\sum_{t=1}^{T} \log p(r_t^p | X, P, r_{1:t-1}^p). \tag{6}$$

The overall learning objective of our proposed method is performed by minimizing:

$$\mathcal{L} = \mathcal{L}_{LLM} + \alpha \mathcal{L}_{CAM}, \tag{7}$$

where α represents the hyperparameter that determine the contribution of $\mathcal{L}_{CAM}$ to the overall loss. We set $\alpha = 0.5$ by default.

3 Experiments

3.1 Dataset and Evaluation Metrics

Dataset. IU X-Ray [4] dataset is a moderately sized MRG dataset, comprising a total of 7,470 pairs of images and 3,955 corresponding reports. We adopt the

Table 2. Model performance with different designed modules in terms of NLG metrics. The AVG refers to the average improvement of all NLG metrics compared with the base model. The highest and the second-highest results are denoted in **bold** and underlines.

Dataset	Base	KIM	CAM	CEM	BLEU-4	ROUGE-L	METEOR	AVG
MIMIC CXR	✓				0.126	0.288	0.154	-
	✓	✓			0.133	0.294	0.164	4.71%
	✓		✓		0.132	0.300	0.165	5.35%
	✓			✓	0.129	0.295	0.162	3.34%
	✓	✓	✓	✓	**0.136**	**0.303**	**0.169**	**7.62%**

same dataset partitioning approach [15, 26], allocating the dataset into training, testing, and validation subsets in a ratio of 7:2:1. MIMIC-CXR [13] is the largest publicly available dataset of chest X-ray images, comprising 377,110 images. We conduct experiments following the MIMIC-CXR's official data split protocol.

Evaluation Metrics. Following the previous works [26, 12], we select two types of evaluation metrics, natural language generation (NLG) and clinical efficacy (CE) metrics, to verify the performance of our proposed method and comparison frameworks.

3.2 Implementation Details

All experiments presented in this paper utilized frontal views from the two datasets. The pre-trained ViT-B/16 is selected as the visual encoder, while the LLaMA3.1-8B is employed as the LLMs-based decoder. The designed KIM consists of a cross-attention module and 3-layer GCN. In the CEM, we choose the top 15 diagnosis reports to serve as external auxiliary information. The proposed model is trained on a single Nvidia A100 80GB GPU, with a batch size of 8. In the MIMIC-CXR dataset, the training epoch is set to 5, whereas in the IU X-Ray dataset, it is configured to 20. The learning rates for the visual encoder and other components are set to 5×10^{-5} and 1×10^{-4}, with a linear decay of 0.9. Specifically, the CLIP and LLMs are frozen, while ViT, BERT, cross-attention, and several project modules need to be trained.

3.3 Main Results

We conduct a comparative analysis of our proposed method against several leading MRG solutions, which are shown in Table 1. Our approach outperforms existing solutions across most indicators in both datasets. In the IU X-Ray dataset, our designed method achieves the best performance on five NLG evaluation metrics, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR. Additionally, it takes 0.385 on ROUGE-L, which is only slightly lower than the 0.390 achieved by BoostRRG [16]. For the MIMIC-CXR dataset, Specifically, our framework achieves 0.414, 0.270, and 0.136 in BLEU-1, BLEU-2, and BLEU-4,

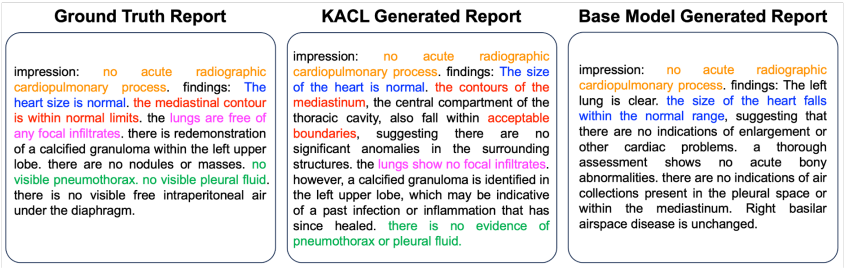

Fig. 3. Qualitative examples of the proposed method and base model. Different colors highlight different medical terms in the reports.

respectively. These results demonstrate an improvement compared to the performance of existing methodologies. However, our approach exhibits slightly lower performance than R2GenGPT [25] in the BLEU-3 metric, primarily because the latter employs more sophisticated fine-tuning techniques and computing resources during training. For the CE metrics, our method significantly outperforms previous SOTA methods in Precision. While the proposed KACL has a lower Recall and F1 Score than PromptMRG [12], this difference is attributed to PromptMRG's use of an external classification model to assign labels to input images.

3.4 Ablation Study

Effectiveness of Proposed Modules. In this part, we evaluate the contribution of the various proposed modules to the overall performance. The detailed results are shown in Table 2. The base model primarily relies on the Visual Encoder and LLMs-based Decoder. The study reveals that each proposed module contributes to improving the model's performance. Compared to the base model, the KIM and CEM provide a notable improvement, with an average NLG increase of 4.71% and 3.34%, respectively. Compared with KIM and CEM, The CAM substantially boosts the base model performance. With the help of three modules, the KACL achieves a relative improvement of 7.62% over the base method.

Qualitative Results. We present a qualitative example that highlights the advantages of KACL compared to the base model, as illustrated in Figure 3. The reports generated by our method capture the majority of key information present in the ground truth report, demonstrating a significant improvement over the base model.

4 Conclusion

In this paper, we propose a knowledge-guided vision-language alignment framework with contrastive learning and LLMs to finish the MRG task. By leveraging

visual features alongside relevant medical knowledge, the powerful LLMs-based decoder can generate more accurate reports. To address challenges in the MRG domain, we develop three modules: KIM, CAM, and CEM, which are intended to integrate medical knowledge, enhance cross-modality feature alignment, and boost diagnostic accuracy, respectively. To demonstrate the performance of the proposed method, we conduct extensive experiments on two datasets, including IU X-Ray and MIMIC-CXR. Experimental results show that our method outperforms existing SOTA methods in most NLG and CE metrics.

Acknowledgments. This study was funded by Science and Technology Development Funds (FDCT) of Macao (0033/2023/RIB2).

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

References

1. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 (2022)
2. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1439–1449 (2020)
3. 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)
4. 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)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Endo, M., Krishnan, R., Krishna, V., Ng, A.Y., Rajpurkar, P.: Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In: Machine Learning for Health. pp. 209–219. PMLR (2021)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
10. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19809–19818 (2023)

11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, 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)
12. 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)
13. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
14. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
15. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., Zhou, L.: Kargen: Knowledge-enhanced automated radiology report generation using large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–392. Springer (2024)
16. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
19. Ramjee, P., Sachdeva, B., Golechha, S., Kulkarni, S., Fulari, G., Murali, K., Jain, M.: Cataractbot: an llm-powered expert-in-the-loop chatbot for cataract patients. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **9**(2), 1–31 (2025)
20. Sha, Y., Meng, W., Luo, G., Zhai, X., Tong, H.H., Wang, Y., Li, K.: Metdit: Transforming and analyzing clinical metabolomics data with convolutional neural networks. Analytical Chemistry **96**(7), 2949–2957 (2024)
21. Sha, Y., Pan, H., Xu, W., Meng, W., Luo, G., Du, X., Zhai, X., Tong, H.H., Shi, C., Li, K.: Mdd-llm: Towards accuracy large language models for major depressive disorder diagnosis. arXiv preprint arXiv:2505.00032 (2025)
22. Sha, Y., Zhang, Q., Zhai, X., Hou, M., Lu, J., Meng, W., Wang, Y., Li, K., Ma, J.: Cervifusionnet: A multi-modal, hybrid cnn-transformer-gru model for enhanced cervical lesion multi-classification. iScience **27**(12) (2024)
23. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)
24. 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)
25. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
26. Yin, H., Zhou, S., Wang, P., Wu, Z., Hao, Y.: Kia: Knowledge-guided implicit vision-language alignment for chest x-ray report generation. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 4096–4108 (2025)