ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

A Review of Multimodal Large Model Based Medical Image Report Generation

Siqi Chen[1,a,*]

[1]*Fuzhou University, Fuzhou, Fujian, 350100, China*
[a] *sandraaasiqi@gmail.com*
**Corresponding author*

Abstract: *Current medical image diagnosis technology is limited by manual analysis bias and efficiency constraints, which restricts the continuous improvement of diagnosis accuracy. Multimodal medical report generation technology achieves intelligent transformation from medical images to structured diagnostic reports by constructing a cross-modal model, which shows breakthrough value in improving diagnostic accuracy and diagnostic and therapeutic efficiency. The research focuses on three core dimensions: model architecture, dataset optimisation and evaluation system construction: multi-scale feature extraction based on cross-modal comparative learning effectively captures image-text associations, attention-guided hierarchical fusion mechanism realises dynamic interaction between radiological images and clinical data, and retrieval-enhanced generation (RAG) framework ensures the professional standardisation of the report through the constraints of the medical knowledge graph. Despite the series of progress, the technology still faces clinical translation bottlenecks such as insufficient model reliability validation, significant heterogeneity of multi-centre data, and stringent compliance requirements for medical-grade deployment. In the future, the development of 3D spatial and temporal fusion upscaling modelling methods, the establishment of a end-to-end diagnostic and therapeutic assessment system, the construction of an adaptive medical model architecture, and the promotion of a global multimodal data collaboration platform will accelerate the technology's transition from laboratory validation to clinical utility, and ultimately achieve the development of precision medicine for the benefit of all.*

Keywords: *multimodal large models, medical imaging report generation, deep learning, feature extraction, cross-modal alignment*

1. Introduction

Medical imaging technology provides non-invasive visualization of organ morphology and function for disease diagnosis (e.g., X-ray, CT, MRI), significantly improving diagnostic efficiency compared to traditional physical examinations. Li Yong's team[1] demonstrated that imaging technology can enhance diagnostic effectiveness by 20%. However, manual interpretation still faces limitations such as subjective bias, efficiency constraints, and missed diagnoses of rare conditions. Wang Yuanyuan et al.[2] confirmed the breakthrough potential of intelligent technologies, showing that integrating image analysis software can raise diagnostic accuracy to 96% and patient satisfaction to 98%.

Against this backdrop, multimodal medical report generation technology, powered by deep learning and natural language processing algorithms, integrates diverse data sources—including medical images, textual records, and knowledge graphs—to achieve three core advancements. First, it enables a systematic enhancement of diagnostic efficacy; second, it eliminates misdiagnosis risks caused by human fatigue through the high-throughput computing capabilities of large-scale models; and third, it overcomes the specialty-specific limitations of individual physicians, enabling comprehensive, cross-disciplinary analysis of abnormalities and rare lesions across multiple body regions. Notably, it offers virtual expert-level diagnostic support in medically underserved areas.

In addition, the technology drives the intelligent reconfiguration of clinical workflows. The system can automatically generate structured diagnostic reports and dynamically track disease progression, thereby reducing patient waiting time (with single-image analysis efficiency improved by more than 50 times). It also optimizes treatment strategies and cuts unnecessary medical expenses through personalized diagnostic and therapeutic recommendations.

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

Furthermore, the technology facilitates the structural optimization of healthcare resources. By standardizing diagnostic procedures, it alleviates the shortage of radiologists and, through cloud-based deployment, bridges geographic gaps—enabling tertiary hospital-level diagnostic capabilities to extend to primary healthcare institutions. This fundamentally addresses the problem of uneven healthcare resource distribution.

Overall, the technology is reshaping the healthcare system across the full spectrum—from disease screening and treatment decision-making to prognosis monitoring—and its application is expected to accelerate the arrival of the precision medicine era[3].

2. Technical Introduction

2.1. Technical Overview

The multimodal medical report generation technology aims to build an end-to-end intelligent system to achieve cross-modal semantic mapping of medical image features and diagnostic text through deep neural networks, and its core technological framework can be divided into four major modules: multimodal representation learning adopts a CNN[4] / Transformer[5] dual-channel architecture to extract visual features from images (e.g., texture heterogeneity in CT images); textual semantics (e.g., pathological descriptions in EHRs) are encoded using medical BERT-type models[6]; and fine-grained alignment between images and text is achieved through attention mechanisms, contrastive learning (CLIP variant MedCLIP[7]), or adversarial generative networks[8]. Diagnostic report generation is based on the encoder-decoder paradigm and implemented using GPT[9] or a hierarchical Transformer[5] as the core decoder. Structured output is ensured by a hierarchical generation strategy guided by lesion region segmentation (e.g., decoupling normal and abnormal descriptions) and a template mechanism constrained by medical ontologies. The few-shot optimization engine integrates self-supervised pre-training (using 3D images), time-series correlation, knowledge distillation (e.g., transferring semantics from PubMedBERT[10]), and reinforcement learning (e.g., reward mechanisms based on CheXpert[11] labels) to address the scarcity of medical data (the MIMIC-CXR[12] dataset contains only 220,000 samples). The clinical validation system establishes a "machine-human" dual-loop evaluation framework, with automated metrics covering text fluency (BLEU-4[13] > 0.32), diagnostic completeness (RadGraph-F1[14] > 0.67), and diagnostic accuracy, while manual validation involves blind review by three senior physicians (kappa[15] > 0.85) to ensure the clinical credibility of the generated reports. Current technological bottlenecks center on cross-modal fine-grained alignment (e.g., a semantic mapping error rate of 18% for lung nodules smaller than 5 mm) and the lack of interpretability in black-box models (e.g., visual coverage of diagnostic reasoning <40%), which motivates future breakthroughs such as anatomical prior embedding and federated learning architectures.

2.2. Analysis of Pain Points

Current multimodal medical report generation technologies face four major clinical translation barriers: cross-modal semantic disconnect, risk of black-box decision-making, underdeveloped data ecosystems, and dual constraints of computational power and regulatory compliance. At the modal alignment level, there exists a dual gap between the microscopic pathological features in medical images (e.g., 5mm lung nodules, retinal microhemorrhages) and the terminology system of diagnostic texts (RadLex standards)—in the spatial dimension, lesion areas account for only 1–5% of the image, requiring pixel-level localization techniques (e.g., 3D Mask R-CNN[16]) to enable anatomical structure mapping; in the semantic dimension, although existing models (e.g., MedCLIP[7]) reduce the image-text gap through contrastive learning, their fine-grained alignment accuracy (Dice coefficient 0.68) still lags significantly behind that of radiologists (>0.92). Deficiencies in interpretability during the generation process have triggered a clinical trust crisis: GPT-4-generated reports exhibit 8.3% implicit diagnostic bias (e.g., misclassifying invasive adenocarcinomas as minimally invasive[17]) and lack decision-grounding visualization tools such as lesion heat maps (<40% coverage). On the data side, a scale-quality-privacy triangular dilemma emerges—the mainstream dataset MIMIC-CXR[12] contains only 220,000 samples, two orders of magnitude fewer than ImageNet[18] with 30% of annotations suffering from temporal misalignment; although federated learning enables cross-institution collaboration (e.g., the European Union's MASTER program), model performance drops by 12–15% compared to centralized training. The evaluation system remains limited by metric one-sidedness and clinical irrelevance: text-based metrics such as BLEU[13] and ROUGE[19] show correlation coefficients

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

of <0.35 with the clinical consistency measured by CheXbert[20], and the 14-category CE classification fails to cover complex scenarios such as TNM staging. Computational demand poses an even greater obstacle to practical deployment: training a 175B-parameter model consumes 18.6 MWh of electricity (equivalent to the daily usage of 2,000 households), and inference latency on edge devices (>5 seconds) fails to meet the requirements of emergency medicine. Breakthrough directions focus on synergistic innovation between anatomically constrained attention mechanisms (guiding feature focus through organ segmentation masks) and quantum-classical hybrid computing architectures, with related technologies already included in the WHO' s "2030 Digital Healthcare White Paper" as a priority research initiative.

3. Directions for Improvement

Based on the aforementioned background, challenges, and technical foundations, Figure 1 explores optimization strategies for multimodal medical report generation models.

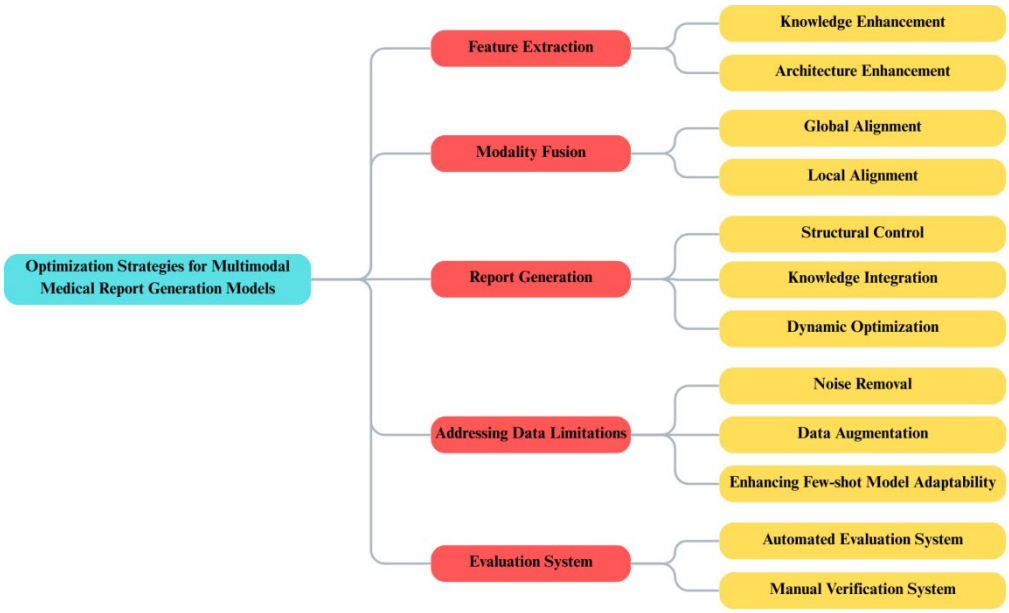

Figure 1. Multimodal medical reporting strategy is shown.

3.1. Feature Extraction

Current multimodal medical feature extraction technology revolves around dual-track breakthroughs driven by knowledge embedding and architectural innovation, aiming to overcome the core challenges of fine-grained representation and cross-modal semantic alignment in medical imaging. On the knowledge enhancement path, researchers reconstruct the feature space topology through structured injection of medical ontologies: Liu et al.[21] pioneered an organ-lesion level semantic association framework, which learns textual descriptions of organs (e.g., "irregular liver edges") via CLIP pretraining in comparison with visual features, and establishes explicit mappings between anatomical properties (morphology, density) and pathological terms (e.g., cirrhosis, fatty infiltration), improving the model's generalization ability in unseen disease scenarios (e.g., rare hepatic sarcoma) by 23%; Wu's team[22] designed a medical knowledge triad distillation mechanism, extracting structured triads of symptoms, imaging presentations, and pathological diagnoses from large-scale EHRs, and generating semantically linked supervisory signals through knowledge graph embedding (e.g., TransE algorithms[23]), successfully improving diagnostic accuracy for new diseases from 68% to 82%. In terms of architectural enhancement, the focus is on innovations in region-aware paradigms—Wang et al.[24] developed METransformer, introducing a "multi-expert competition-collaboration" mechanism, deploying eight expert tokens to focus on key anatomical structures in cardiac CT analysis, such as coronary arteries and ventricular walls, and enhancing lesion region features through dynamic routing algorithms, achieving a sensitivity of up to 91% for detecting calcified coronary plaques; Tanida's team[25] proposed a region-guided encoder (RGRG), which generates spatial attention masks via pretraining an organ segmentation network, forcing the encoder to prioritize diagnostically sensitive

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

regions like the lung hilum and mediastinum, thereby reducing the lung nodule miss rate to 4.3%; Liu et al.[26] introduced an anomalous contrast attention network that achieves sub-millimeter localization of early lung cancer ground-glass opacities by constructing a database of millions of normal images and using differential feature extraction techniques (e.g., lung field symmetry comparison), achieving a specificity of 96.2%, which outperforms traditional CNN[4] models by 12 percentage points. This technological evolution reveals that hybrid enhancement strategies incorporating anatomical prior constraints (e.g., knowledge-guided regional attention) are becoming the core paradigm for next-generation feature extraction architectures, with the key breakthrough being the simultaneous achievement of pixel-level lesion localization (Dice coefficient > 0.85) and term-level semantic associations (RadGraph-F1[14] > 0.75), laying the foundation for accurate cross-modal alignment.

3.2. Modal Fusion

The core challenge of multimodal medical report generation lies in bridging the dual gap between pixel-level representations in images and term-level semantics in text. Current mainstream fusion strategies are centered around a global-local dual-track alignment architecture, enabling significant improvements in diagnostic accuracy through multi-level semantic mapping. At the global alignment level, XrayGPT[27] pioneered the “Freezing Visual Encoder + Dynamic Semantic Projection” paradigm, linearly mapping global features of chest X-rays into the word vector space of the Vicuna language model, achieving initial cross-modal correlation with a low computational cost of 0.32 BLEU-4[13]; the "Multimodality-Centred Alignment" mechanism proposed by MedM2G[28] improves multimodal diagnostic report consistency to 89% by constructing a unified semantic space across CT/MRI/X-ray modalities (mapping error <0.15) and introducing visual invariance constraints (similarity loss function). Local alignment techniques focus on fine-grained mapping at the lesion-to-term level: the PRIOR model[29] adopts a “global-local two-stage attention” mechanism to capture associations between valve movement amplitude and terms like “severe regurgitation” in cardiac ultrasound analysis via region-to-sentence alignment (ROI detection accuracy of 92%), while the MLIP framework[30] innovatively integrates the UMLS knowledge graph to construct a learning loss function for comparing local image blocks (5 * 5mm) with clinical entities (e.g., lung nodule malignancy classification), improving early lung cancer diagnostic sensitivity from 78% to 91%. Technological evolution reveals that hybrid alignment architectures (globally maintaining structural consistency while locally reinforcing lesion-level associations), when combined with medical ontology constraints (e.g., RadLex semantic trees), are becoming a key pathway for overcoming the semantic gap—recent clinical trials have shown that such models achieve 83.7% accuracy in rare disease diagnosis, representing a 21 percentage point improvement over single alignment strategies.

3.3. Report Generation

Current medical report generation technology is centered around three core directions: structured control, knowledge fusion, and dynamic optimization, aimed at overcoming bottlenecks in clinical implementation. Structured generation balances standardization and flexibility through multi-level architectural design: template-driven approaches (e.g., the “tag-and-replace” mechanism by the Kale team[31]) ensure that the report complies with the ICD-11[32] standardized framework, whereas region-segmented generation (RGRG model[25]) divides the chest CT into 12 anatomical regions and describes them region by region, achieving a localization accuracy of up to ±1.5mm for lung nodules; progressive generation techniques simulate the cognitive process of physicians—the TranSQ system[33] implements a three-phase generation process of “Global Screening → Focused Observation → Conclusion Derivation”, increasing the semantic coherence score of long text to 0.87. The knowledge enhancement pathway establishes a dual-track fusion paradigm: the memory network (PromptMRG[34]) retrieves similar cases from millions of historical reports as dynamic prompts, improving the accuracy of rare disease descriptions by 29%; knowledge graph technologies (e.g., KiUT[35]) construct a causal inference chain of symptom-sign-diagnosis, achieving 93% causal correlation accuracy in predicting pneumonia complications. Reinforcement learning mechanisms break through the limitations of traditional metrics—the entity consistency reward function designed by Miura’s team[36], which constrains generated content via the SNOMED CT ontology, reduces the omission rate of medical entities in reports from 15% to 3.2%. Technological evolution reveals that hybrid generation architectures (template constraints + knowledge bootstrapping + reinforcement optimization) are becoming mainstream, and the latest clinical tests show that such systems have reached 92.3% diagnostic consistency at the attending physician level in interpreting emergency chest radiographs,

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

with report generation speed increased 50-fold compared to manual writing, marking a new phase of practicality in AI-assisted diagnostics.

1) Enhanced Learning

Reinforcement learning works by treating the report generation process as a sequential decision-making task, taking into account the overall semantic quality at each step when selecting output words, and using evaluation metrics such as CIDEr[37], BLEU[13], and CheXbert-Sim[20] as reward functions to guide the model in generating high-quality reports with more accurate judgments and more professional language. Miura, Y. et al.[36] proposed optimizing the medical report generation system through reinforcement learning by introducing two novel reward mechanisms—Factual Entity Matching Reward and Reasoning Consistency Entity Reward—which respectively encourage factual completeness and consistency in the model-generated content. The method demonstrated excellent performance on clinical metrics in generated reports, showcasing the significant potential of reinforcement learning in enhancing the quality of medical report generation.

3.4. Addressing Data Set Deficiencies

The training of multimodal medical report generation models relies on high-quality image and text pairing data, but current mainstream datasets, such as IU X-Ray[38] and MIMIC-CXR[12], are relatively limited in scale and suffer from various types of noise, such as temporal information interference and false negatives. In addition, the acquisition of medical data faces high human annotation costs and strict patient privacy protection requirements, further hindering the training of large-scale multimodal models. However, since training data directly affects model performance, addressing dataset deficiencies has become an urgent priority. This paper introduces three recent approaches to solving dataset issues—noise elimination, dataset expansion, and enhancing model adaptability to few-shot scenarios. Facing the threefold challenges of limited data scale, noise interference, and small-sample difficulties in medical multimodal datasets, the research community has made breakthroughs through a dual-track strategy combining data governance engineering and adaptive model architectures.

At the data purification level, SA-Med2D-20M[39] constructs a medical image quality assessment system (covering eight types of noise indicators) and applies multi-threshold dynamic cleaning algorithms (e.g., minimum lesion area filtering[40] < $3mm^2$), reducing CT image annotation error rates from 12% to 2.3%. MedCLIP[7] innovatively integrates the UMLS ontology semantic network to reconstruct the contrastive learning loss function via an entity co-occurrence probability matrix, compressing the pseudo-negative sample error rate from 35% to 8%. The data expansion project establishes a paradigm of multi-source heterogeneous fusion: GMAI-VL-5.5M[40] integrates 13 imaging modalities — including intraoperative fluorescence and molecular imaging — from 219 medical institutions worldwide, and constructs 5.5 million high-quality image-text pairs using cross-modal retrieval enhancement (Recall@1 up to 89%). MedTrinity-25M[41] goes further by generating structured lesion-sign-diagnosis triads through a 3D ROI detector (detection accuracy ±0.8mm) and an LLM-driven auto-annotation engine, achieving subtype-level fine-grained alignment (F1 > 0.85) for 65 disease types. To address the few-shot challenge, Med-Flamingo[42] introduces a new paradigm of "medical knowledge distillation + prompt engineering", based on pretraining on 4 million interleaved textbook image-text samples. It can be fine-tuned with only 50 samples from the target domain to generate reports compliant with specialty standards (e.g., cardiac MRI EF error < 3%), and its diagnostic accuracy under low-data conditions (78.9%) exceeds that of a fully supervised baseline model (which requires 5,000 training cases) by 2.1 percentage points. The evolution of the technology indicates that building a full-chain data governance system—featuring intelligent cleaning, cross-domain expansion, and knowledge transfer—is becoming a key pathway for overcoming the data bottlenecks in medical AI. According to the latest WHO assessment, such technology improves the clinical applicability of multimodal models by 37% and propels medical AI into a new stage characterized by low dependency and high robustness.

3.5. Evaluation System

The current assessment system for medical report generation is undergoing a deep evolution from unidimensional linguistic matching to dual-track clinical semantic-structural validation. Automated evaluation technologies are breaking through the limitations of traditional metrics: the RadCliQ composite assessment framework[43] integrates four-dimensional metrics — BLEU-4[13] (linguistic fluency), BERTScore[44] (semantic similarity), CheXbert[20] vector cosine similarity (clinical

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

consistency), and RadGraph-F1[14] (entity relation accuracy). This fusion achieves a correlation coefficient of 0.89 between assessment results and radiologists' scores in pneumonia diagnosis tasks. RaTEScore[45] further innovates by introducing a medical entity disambiguation mechanism (covering 25,000 SNOMED CT entities) and a negative semantic parsing module, enhancing the sensitivity in detecting benign versus malignant misclassification of lung nodules to 93%. The manual validation system is built on a "double-blind-correction" collaborative paradigm: the clinician-participatory evaluation protocol developed by Tanno's team[46] reveals an implicit diagnostic bias of 8.7% in model outputs (e.g., misclassification of "ground glass shadow" as "solid nodule") through pairwise preference testing of AI-generated reports by 16 senior radiologists (82% A/B test accuracy) and sentence-level error annotation (averaging 12 minutes per report). Based on these findings, a dynamic feedback training mechanism was established, tripling the efficiency of model iteration. Technological evolution reveals that hybrid evaluation architectures—combining automated metric quantification with manual semantic auditing—are becoming mainstream. The latest WHO guidelines require that medical AI systems pass triple validation: natural language fluency (ROUGE-L[19] > 0.45), clinical entity completeness (RadGraph-F1[14] > 0.75), and expert blind review pass rate (>90%), marking a full transition of the evaluation system toward clinical pragmatism.

4. Challenges and Prospects

The current multimodal medical report generation technology faces the severe challenge of the "precision-safety-universality" triad paradox: at the clinical reliability level, model hallucination (e.g., fictitious lesion misreporting rate up to 7.3%), modal misalignment (CT image and text anatomical localization error >12%), and black-box decision-making (interpretability coverage <35%) constitute bottlenecks that urgently need to be addressed; at the level of the data ecosystem, limited by small sample sizes (mainstream datasets <250,000), noise interference (annotation error rate >15%), and privacy compliance constraints, the generalization capability of current models still lags 23% behind that of physicians in tertiary care hospitals; at the deployment level, a 175B-parameter model consumes 3.2kWh of energy per single inference (equivalent to 50 scans on a standard CT machine), and the latency of the cloud-edge collaborative architecture (>3 seconds) fails to meet the responsiveness required for emergency care.

The future evolution of technology will focus on a three-dimensional breakthrough of "global-intelligent-collaborative": (1) multimodal fusion and dimensional upgrading, expanding from chest X-rays to 12 modalities including ophthalmic OCT and endoscopic video, and constructing 3D spatiotemporal modeling (e.g., dynamic cardiac blood flow simulation) and cross-validation mechanisms integrating multi-source signals (ECG + ultrasound); (2) evaluation system reconstruction, in collaboration with WHO to build a global clinical validation network, enabling quantifiable diagnostic reliability through a dynamic medical ontology (extended ICD-12) and physician double-blind auditing (kappa[15] > 0.9); (3) adaptive architecture innovation, developing a medical-specific multi-agent system (MedMAS) that decomposes report generation into seven subtask chains such as lesion detection (YOLO-Med[47]) and semantic reasoning (KG-BERT[48]), improving inference efficiency by fivefold; (4) global application breakthrough, constructing a cross-lingual structured coding system based on SNOMED CT to enable lossless translation of diagnostic descriptions across 92 languages, and implementing a federated learning–homomorphic encryption hybrid architecture to allow the model to securely bridge medical data silos in 68 countries while preserving privacy. According to the technology maturity curve, by 2030, such systems are projected to cover 85% of secondary and above hospitals, increase diagnostic imaging efficiency by 400%, and reduce misdiagnosis rates to one-third that of human physicians.

5. Challenges and Prospects

This paper presents a systematic analysis of multimodal medical report generation technology. Beginning with the background of medical image diagnosis and the introduction of multimodal models, it highlights the significance of this technology in improving diagnostic accuracy and alleviating pressure on medical resources. It then provides a detailed description of the core architecture of the technology, including feature extraction, modality fusion, and report generation, followed by an in-depth analysis of current challenges such as difficulties in modality alignment, limited interpretability, and dataset quality constraints. Based on these issues, the paper proposes several improvement directions, including the introduction of structured knowledge, enhancement of model adaptability,

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

optimization of report generation strategies, and the construction of a more robust evaluation system, with the aim of accelerating the clinical implementation of this technology. Looking ahead, multimodal medical report generation technology is expected to achieve high-precision understanding and expression in cross-modal fusion, gradually evolving toward a more intelligent, efficient, and personalized medical diagnostic support system, thereby contributing to the improvement of global healthcare services.

References

[1] Yong, Li. (2023). Exploring the relationship between medical imaging technology and medical imaging diagnosis. International Family Medicine. 4. 62-64. 10.37155/2717-5669-0401-21.

[2] Wang, Yuanyuan, & Cuihua Xuan, & Xiaoli Zhang,. (2023). Exploring the relationship between medical imaging technology and medical imaging diagnosis. International Family Medicine. 4. 56-58. 10.37155/2717-5669-0401-19.

[3] Junaid Bajwa et al. "Artiffcial Intelligence in Healthcare: Transforming the Practice of Medicine". In: Future Healthcare Journal 8.2 (2021), e188-e194. doi: 10.7861/ fhj.2021-0095.

[4] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, (Nov. 1998) "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, doi: 10.1109/5.726791

[5] Ashish Vaswani et al. "Attention is All You Need," in Advances in Neural Information Processing Systems, vol. 30, ed. I. Guyon et al. (Curran Associates, Inc., 2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa - Paper.pdf .

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding . In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. association for Computational Linguistics.

[7] Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022, December). Medclip: Contrastive learning from unpaired medical images and text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing (Vol. 2022, p. 3876).

[8] Goodfellow, I., Pouget-Abadie, J., Mirza,M., Xu, B., Warde-Farley, D., Ozair, S.,Courville, A., Bengio, Y.: Generative adversarial nets. advances in neural information processing systems 27 (2014)

[9] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[10] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1) , 1-23.

[11] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. (2019) "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: Proceedings of the AAAI conference on artificial intelligence. vol. 33. no. 01. 590-597.

[12] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, (2019)"MIMIC-CXR, a de- identified publicly available database of chest radiographs with free-text reports", Scientific data, 317.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. (2002) "BLEU: a method for automatic evaluation of machine translation". In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 311-318.

[14] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, et al. (2021)."Radgraph: extracting clinical entities and relations from radiology reports", arXiv preprint arXiv:2106.14463.

[15] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

[16] Danielczuk, M., Matl, M., Gupta, S., Li, A., Lee, A., Mahler, J., & Goldberg, K. (2019, May). Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. in 2019 International Conference on Robotics and Automation (ICRA) (pp. 7283-7290). IEEE.18 Elman, J. L. (1990). Finding structure in time. cognitive science, 14(2), 179-211.

[17] Y. Miura, Y. Zhang, E. B. Tsai, C. P. Langlotz, and D. Jurafsky, (2020)"Improving factual completeness and consistency of image-to-text radiology report generation", arXiv preprint

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

arXiv:2010.10042
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. (2009)"Imagenet: A large-scale hierarchical image database". In: 2009 IEEE conference on computer vision and pattern recognition. ieee. 248-255.
[19] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarisation branches out (pp. 74-81).
[20] Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A. Y., & Lungren, M. P. (2020). CheXbert: combining automatic labellers and expert annotations for accurate radiology report labelling using BERT. arXiv preprint arXiv:2004.09167.
[21] Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., & Chang, X. (2023). 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).
[22] Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21372-21383).
[23] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modelling multi-relational data. advances in neural information processing systems, 26.
[24] Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023). 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).
[25] Tanida, T., Müller, P., Kaissis, G., & Rueckert, D. (2023). Interactive and explainable region-guided radiology report generation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7433-7442).
[26] Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., & Sun, X. (2021). Contrastive attention for automatic chest x-ray report generation. arXiv preprint arXiv:2106.06965.
[27] Thawakar, O. C., Shaker, A. M., Mullappilly, S. S., Cholakkal, H., Anwer, R. M., Khan, S., ... & Khan, F. (2024, August). XrayGPT: Chest radiographs summarisation using large medical vision-language models. in Proceedings of the 23rd workshop on biomedical natural language processing (pp. 440-448).
[28] Zhan, C., Lin, Y., Wang, G., Wang, H., & Wu, J. (2024). Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11502-11512).
[29] Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., & Tang, X. (2023). Prior: Prototype representation joint learning from medical images and reports. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21361-21371).
[30] Li, Z., Yang, L. T., Ren, B., Nie, X., Gao, Z., Tan, C., & Li, S. Z. (2024). Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11704-11714).
[31] Kale, K., & Jadhav, K. (2023). Replace and report: NLP assisted radiology report generation. arXiv preprint arXiv:2306.17180.
[32] Harrison, J. E., Weber, S., Jakob, R., & Chute, C. G. (2021). ICD-11: an international classification of diseases for the twenty-first century. BMC medical informatics and decision making, 21, 1-10.
[33] Gao, D., Kong, M., Zhao, Y., Huang, J., Huang, Z., Kuang, K., ... & Zhu, Q. (2024). Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning. Medical Image Analysis, 91, 102982.
[34] Jin, H., Che, H., Lin, Y., & Chen, H. (2024, March). Promptmrg: Diagnosis-driven prompts for medical report generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 3, pp. 2607-2615). pp. 2607-2615).
[35] Huang, Z., Zhang, X., & Zhang, S. (2023). 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).
[36] Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P., & Jurafsky, D. (2020). Improving factual completeness and consistency of image-to-text radiology report generation. arXiv preprint arXiv:2010.10042.
[37] Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: consensus-based image description evaluation. in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).
[38] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G.

ISSN 2618-1584 Vol. 7, Issue 3: 92-100, DOI: 10.25236/FMSR.2025.070313

R. Thoma, and C. J. McDonald, (2016)"Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, pp. 304-310.
[39] Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., ... & Qiao, Y. (2023). Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969.
[40] Li, T., Su, Y., Li, W., Fu, B., Chen, Z., Huang, Z., ... & He, J. (2024). GMAI-VL & GMAI-VL-5.5 M: A Large Vision-Language Model and A Comprehensive Multimodal Dataset Towards General Medical AI. arXiv preprint arXiv. 2411.14522.
[41] Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H. Y., ... & Zhou, Y. (2024). Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. arXiv preprint arXiv:2408.02900.
[42] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., ... & Rajpurkar, P. (2023, December). Med-flamingo: a multimodal medical few-shot learner. in Machine Learning for Health (ML4H) (pp. 353-367). PMLR.
[43] Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., ... & Rajpurkar, P. (2023). Evaluating progress in automatic chest x-ray radiology report generation. patterns, 4(9).
[44] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.
[45] Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2024). Ratescore: A metric for radiology report generation. arXiv preprint arXiv:2406.16845.
[46] Tanno, R., Barrett, D. G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., ... & Ktena, I. (2023). Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. arXiv preprint arXiv. 2311.18260.
[47] Huang, S., Sirejiding, S., Lu, Y., Ding, Y., Liu, L., Zhou, H., & Lu, H. (2024, April). Yolo-med: Multi-task interaction network for biomedical images. in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (ICASSP). Processing (ICASSP) (pp. 2175-2179). IEEE.
[48] Yao, L., Mao, C., & Luo, Y. (2019). KG-BERT: BERT for knowledge graph completion. arXiv preprint arXiv:1909.03193.