Recent Advances In Multimodal Transformers For Medical Image And Clinical Text Analysis

Sahil Sharma[1], Dr. Rajinder Kumar[2], Dr. Tarandeep Singh Walia[3], Deepender[4]

[1]*Research Scholar, Department of CSE, Guru Kashi University, Bathinda, Punjab, India*
[2]*Associate Professor, Faculty of Computing, Guru Kashi University, Punjab, India*
[3,4]*School of Computer Applications, Lovely Professional University, Punjab, India*

Abstract

Multimodal transformers, the latest AI models, can understand medical images along with clinical text like reports and patient notes. These models help doctors diagnose diseases more accurately and quickly. This review explains how these models are built and the data they learn from. Here, the challenges such as limited data, privacy concerns and the need for better explainability, are also explained in brief. Finally, the review also highlights future goals such as stronger and better foundation models and better standards for evaluating these systems.

Keywords: Multimodal transformers, AI, Medical images

1. Introduction

In the field of medical diagnosis, the doctors often rely on different types of information like medical images, X-rays, CT scans and pathology slides. Along with it also takes into account written records such as clinical notes and structured data from lab tests and patient historical records. When these different sources are combined, they help provide a more complete overview of a patient's condition and can lead to more accurate decisions than relying on a single source.

Transformer based models proved to be strong solutions for bringing these data together. Their self attention and cross modal attention mechanisms help the model to learn how information acquired from images, texts and clinical data relate to each other. As a result, they support tasks like retrieving related data across different formats, generating medical reports and assisting with diagnosis.

In recent years, the field has shifted away from older CNN and RNN combinations toward unified multimodal transformer models that can be pre-trained and fine-tuned specifically for healthcare. This review examines the major developments from 2019 to 2024, covering techniques, datasets, performance trends, current challenges and opportunities for future innovation.

2. Background

Transformers were first created for natural language processing, but they have since been adapted very effectively for combining visual and textual information. Early vision–language models like ViLBERT and VisualBERT introduced two different design strategies. One, using separate streams for image and text features and another using a single shared stream. Both allowed the model to learn how images and text relate to one another through mechanisms such as co-attention and shared embeddings. In the medical field, the text-processing components are often replaced with biomedical language models like BioBERT, which are trained on clinical documents to better understand medical terminology. These early innovations provided the foundation for healthcare-focused multimodal transformers that take advantage of naturally paired datasets such as chest X-ray images and their corresponding radiology reports, to learn clinically meaningful representations.

3. Technical Aspects

Multimodal Transformers build on the standard self-attention mechanism by allowing information to flow between

different types of data—most commonly, medical images and clinical text. In these models, both images and text are converted into sequences of vector embeddings with the same dimensional size. For example:

- V represents visual tokens, which may come from a CNN encoder or image patches.
- T represents textual tokens extracted from reports or clinical notes.

Once both are in this shared embedding space, the Transformer can perform cross-modal attention. In one direction, **visual-to-text attention** uses image features to actively query the textual information. Here visual tokens take the role of queries, while the text tokens serve as keys and values.

$$Q_v = VW_Q^V, K_T = TW_K^T, V_T = TW_V^T$$
$$Attention_{V\to T} = softmax\left(\frac{Q_V K_T^T}{\sqrt{d}}\right)V_t$$

This equation describes **visual-to-text attention**, where each image region acts as a query (Q_V) to find relevant information in the text, which is represented by keys (K_T) and values (V_T). This model computes similarity score between image queries and text keys. Then scales them and applies a softmax to acquire attention weights. Then, these weights are used to combine the text values and produce a text informed representation of each image region. Basically, it allows the model to highlight those parts of the text that are most relevant to what it "sees" in the image. This further improves tasks like report generation and interpretability. The system's interpretation of disease indicators is enhanced by the semantic connections made between these picture regions and clinical terminology.

In case of **text-to-visual attention model, (**Report queries to Image regions) the textual information helps the model to locate relevant visual evidence within the image provided.

$$Q_T = TW_Q^T, K_V = VW_K^V, V_V = VW_V^V$$
$$Attention_{T\to V} = softmax\left(\frac{Q_T K_V^T}{\sqrt{d}}\right)V_V$$

The equation above represents **text to visual attention**, where each word in the text queries (Q_T), the image regions (K_V and V_V) are used to find relevant visual information. The model works by calculating similarity score between text queries and image keys. Then scales them and applies softmax to get the attention weights. These weights are then used to combine the image values to produce visual representation guided by the text provided. So basically, it helps the model focus on the parts of the image that correspond to the text. This process helps enabling tasks which are crucial for clinically meaningful tasks like localizing abnormalities, generating accurate medical reports and support more transparent diagnostic decisions.

Next in the **joint fusion layer**, once both attention directions have been computed, their outputs are then combined and mapped into a shared embedded space.

$$H = Concat[Attention_{V\to T}, Attention_{T\to V}]W_H$$

The two attention outputs (One showing how image regions converted to text ($Attention_{V\to T}$) and the other showing how text mapped to image regions ($Attention_{T\to V}$)), are merged and then projected using a weight matrix (W_H). This produces a fused, context-aware feature that further captures the interactions between text and image which can then be used for various tasks like disease classification or generating clinical reports.

Ultimately, this cross-modal attention forms the basis of these transformation systems, that can align visual findings with medical language at a deep semantic level.

4. Multimodal Strategies in Medical Domains

Medical multimodal learning typically uses one of three fusion strategies mentioned below:

Early Fusion: This approach works well when data sources are closely aligned. When data sources are closely linked, this strategy performs well. Before proceeding into shared Transformer levels, features from each modality are concatenated. But when particular techniques are absent, performance may decline.

Late Fusion: The final decisions are aggregated using ensemble rules or classifier heads after each modality is analyzed separately. This approach provides little interaction between modalities, although still being more resilient to missing data..

Cross-Modal (Co-Attention) Fusion: In this modal, Attention mechanisms allow separate visual and textual streams to communicate. This enables fine-grained alignment between image regions and clinical report tokens. This has become the dominant solution in radiology applications [1], [7].
Recent advances point to single stream integrated transformers, in which image patches and text tokens are processed together within the same architecture. These models simplify the pipeline and enable more complex multimodal reasoning.

B. Pretraining Objectives
Medical multimodal models often rely on many pre-training objectives to learn from huge, but imperfect clinical datasets. These models are-

Masked Language Modeling (MLM) is applied to clinical text using domain-specific encoders like BioBERT or ClinicalBERT. This allows greater understanding of medical language [3].

Masked Image Modeling trains the visual encoder to reconstruct missing image patches or parts, which helps in capturing anatomical details.

Contrastive learning links paired images and reports by pulling related representations closer in embedding space. This approach works especially well when dense annotations are limited, as demonstrated in models like ConVIRT and MedCLIP [6], [10].

Knowledge-aware objectives incorporate medical ontologies and entity alignment to ensure that the learned features remain clinically consistent, reducing errors such as mismatched disease labels (e.g. GLoRIA and related extensions) [7], [11].

Together, these strategies improve data efficiency, strengthen robustness across modalities and support better generalization in real clinical environments.

5. Datasets and Benchmarks (2019–2024)

Advances in multimodal medical AI have largely been enabled by the availability of large-scale image report datasets. Some of these are-

CheXpert provides chest X-ray images with pathology labels that account for diagnostic uncertainty, along with radiologist comparison studies [4].

MIMIC-CXR offers a substantial collection of de-identified chest radiographs paired with free-text radiology reports, making it a key resource for tasks such as contrastive pretraining, report generation and zero or few shot learning [5].

Additional data sources include institutional radiology archives, multi-view imaging datasets and specialized pathology collections that combine microscopic images with descriptive reports.
These datasets support common benchmarks like pathology classification, automated report generation and few shot diagnostic evaluation. However, variations in dataset preprocessing, label extraction methods, train-test splits and evaluation metrics often make it difficult to directly compare results across different studies.

6. Key Studies

A. Contrastive Pretraining & Representation Learning

ConVIRT (2020) proposed contrastive pretraining on chest X-ray image-report pairs, significantly improving representation quality and enabling downstream classification and retrieval even with limited labels [6].

MedCLIP (2022) extended contrastive vision-language learning to unpaired medical images and text by decoupling image-text pairs and applying semantic matching losses to reduce false negatives. This achieved strong zero-shot performance and surpassed supervised baselines using far less data [10].

B. Global-Local & Knowledge Enhanced Representations

GLoRIA (ICCV 2021) introduced a global local attention framework that contrasts image sub-regions with report words, achieving label-efficient classification, retrieval and segmentation across tasks [7].

Other works incorporate medical ontologies or knowledge-guided alignment to improve semantic grounding and robustness to dataset shifts [11].

C. Radiology Report Generation

Several Transformer-based encoder–decoder systems generate free-text reports from imaging data, using cross-modal memory, attention-driven fusion and multimodal embeddings to generate clinically coherent reports [8-9], [16].

D. Multi-Modal Fusion

Emerging systems incorporate not only imaging and radiology text, but also structured EHR parameters and clinical notes, aiming for prognosis, triage and comprehensive decision support. These models often rely on cross-modal transformers or adapter-based fusion networks. They show early promise, but require harmonized, high-quality EHR data [12], [18].

7. Comparative Methodology Overview

The table mentioned below summarizes some of the recent work in medical AI that combines different types of data like - images, text reports and patient records. It covers studies from 2020 to 2024 showing the kinds of tasks these models are used for such as classifying diseases, retrieving relevant information, generating reports and predicting patient outcomes. It also explains the techniques they use to combine or pretrain on multiple data sources like contrastive learning, attention mechanisms and knowledge-guided methods. These Studies show that these models can learn efficiently with less data and produce more accurate or relevant outcomes. These models also increase performance across varied datasets, making them applicable to real-world clinical applications.

Reference/ Year	Modal	Main Task(s)	Fusion/ Pretraining Strategy	Key Contributions
ConVIRT (2020)	CXR+ Report text	Representational learning, classification	Contrast image report pretraining	Data efficient embeddings, stronger transfer with limited labels [6]
GLoRIA (2021)	CXR + Report text	Classification, retrieval, segmentation	Global-local attention + contrastive	Label-efficient learning, good region-word alignment [7]
MedCLIP (2022)	Medical images + unpaired text	Zero-shot and supervised classification	Semantic aware contrastive on unpaired data	Shows strong zero-shot results, also reduces data dependency [10]
Knowledge-enhanced VL models (2023)	Imaging + report text + medical	Diagnosis, grounding, retrieval	KG-guided multimodal pretraining	Better semantic consistency, robustness across datasets [11]

	ontology			
Transformer Report Generators (2021 - 2023)	CXR to text report	Report generation	Cross-modal encoder and decoder	More fluent & clinically relevant reports [8] [9]
Multimodal EHR-fusion models (2022-2024)	Imaging + text + structured EHR	Prognosis/triage/ diagnosis	Multi-input fusion networks or adapters	Early gains in prognosis, highlight need for harmonized EHR data [12], [18]

To assist to see the differences and trends across various multimodal medical AI models, the figure below presents a comparison. It describes the modalities used, essential tasks, fusion techniques and key findings. This picture helps to rapidly understand how each strategy integrates data and the benefits it provides in clinical applications.

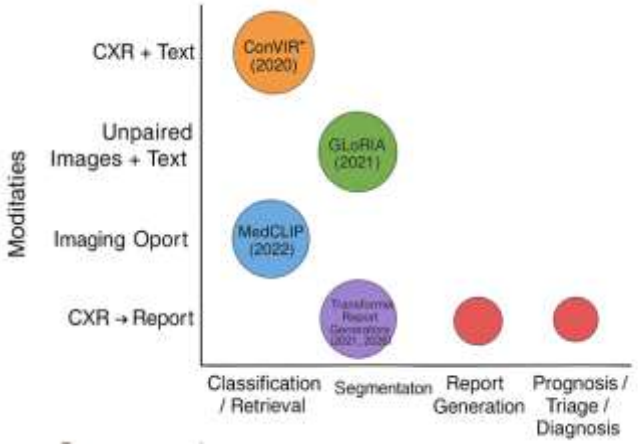

Figure1. Trends across these multimodal medical AI models

8. Benefits

Recent breakthroughs in multimodal medical AI have resulted in significant gains in diagnosis accuracy, generalization and clinical utility. Contrastive and knowledge-enhanced models such as MedCLIP and GLoRIA, performed well in both zero-shot and few-shot chest radiograph categorization. It not only achieves the accuracy of fully supervised algorithms, but also significantly decreases the need for labeled data using effective pretraining strategies.

Transformer based report generation systems produce more coherent and clinically relevant reports than previous CNN+RNN models. This advancements increase their practical utility in radiology operations. Global-local attention mechanisms allowed improved alignment between picture regions and specific clinical terms, improving interpretability and increasing clinician trust. Additionally, developing multimodal fusion frameworks that merge images, clinical text and structured EHR data lay the groundwork for more complete and context-aware decision support systems.

Although these approaches are still in their early phases of development, they have proven to have significant consequences in the relevant fields.

9. Challenges and Limitations

Although these promising results, multimodal medical AI still faces numerous significant hurdles. Radiology reports range widely in style, language and completeness, which can contribute noise and discrepancies. All these challenges make the model training and evaluation challenging [4] [5]. Clinical data are often incomplete, some patients may lack reports, EHR entries or any other structured lab measurements, which limits the effectiveness of models that rely on.

Interpretability is another concern. Attention maps and embeddings are not always intuitive for clinicians and without expert in the loop validation, trust and adoption remains low. Privacy and regulatory issues make sharing data across institutions challenging and can reduce generalizability. On top of that, inconsistencies in dataset preparation, label extraction and evaluation standards across studies make fair comparisons and reproducibility difficult.

10. Future Directions

Building on current advances while addressing existing limitations, several research directions show strong potential for the future of multimodal medical AI. Developing large-scale multimodal foundation models that are pretrained on a variety of imaging modalities-such as X-ray, CT, MRI etc. and pathology, alongside clinical reports and structured EHR data could enable few-shot or zero-shot adaptation across many diagnostic tasks. Models that are uncertainty aware and causally informed could explicitly handle missing modalities, ambiguous reports and conflicting evidence. Establishing standardized benchmarks and evaluation protocols, would greatly enhance reproducibility and fair comparison across studies. In conclusion, clinical-focused frameworks that include attention-based training, medical terminologies, prominence maps and human-in-the-loop validation are critical for building trust and increasing adoption in real-world healthcare settings. Some of the solutions that can be worked upon are mentioned as-

- Large-scale multimodal foundation models
- Privacy-preserving federated and decentralized training
- Uncertainty-aware and causal fusion models
- Standardized benchmarks and evaluation protocols
- Clinician-centered explainability frameworks

11. Conclusion

Multimodal transformer models have great potential to improve how doctors diagnose and treat patients by combining medical images, clinical notes and structured data. The area has rapidly progressed from early contrastive learning approaches like ConVIRT to more advanced knowledge-based systems like GLoRIA and MedCLIP. These approaches have demonstrated significant improvements in tasks such as disease categorization, image-text retrieval and report production. However, issues like inconsistent data quality, privacy problems, missing information, restricted interpretability and difficulties in duplicating results persist. Overcoming these issues will be essential to make multimodal AI safe, reliable and useful in real clinical practice. In future, we can work on large multimodal foundation models, privacy-preserving federated learning and standardized benchmarks to become common in medical AI research.

References

1. J. Lu, D. Batra, D. Parikh and S. Lee, “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks,” in NeurIPS, 2019.
2. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh and K.-W. Chang, “VisualBERT: A Simple and Performant Baseline for Vision and Language,” arXiv preprint arXiv:1908.03557, 2019.
3. J. Lee, W. Yoon, S. Kim et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
4. J. Irvin et al., “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” arXiv:1901.07031, 2019.
5. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific Data, 6, 317, 2019.
6. Y. Zhang, H. Jiang, Y. Miura, C. D. Manning and C. P. Langlotz, “Contrastive Learning of Medical Visual Representations from Paired Images and Text,” arXiv:2010.00747, 2020.
7. S.-C. Huang, L. Shen, M. P. Lungren and S. Yeung, “GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition,” in ICCV, 2021, pp. 3942–3951.
8. Z. Chen, K. Lin and others, “Generating Radiology Reports via Memory-Driven Transformer,” in EMNLP, 2020.
9. Z. Chen, J. Liu and colleagues, “Cross-Modal Memory Networks for Radiology Report Generation,” in ACL, 2021.
10. Z. Wang, Z. Wu, D. Agarwal and J. Sun, “MedCLIP: Contrastive Learning from Unpaired Medical Images and Text,” in

EMNLP, 2022, pp. 3876–3887.
11. C. Wu et al., “MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-Ray Diagnosis,” in ICCV, 2023.
12. X. Hou et al., “MKCL: Medical Knowledge with Contrastive Learning Model for Radiology Report Generation,” J. Biomed. Inform., 2023.
13. H. Lai et al., “CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification,” in CVPR, 2024.
14. S. Nerella et al., “Transformers and large language models in healthcare: a survey,” arXiv preprint arXiv:2402.01234, 2024.
15. T. Koleilat, H. Asgariandehkordi, H. Rivaz and Y. Xiao, “MedCLIP-SAMv2: Towards Universal Text-Driven Medical Image Segmentation,” arXiv:2409.19483, 2024.
16. U. Rahman, R. Imam, D. Mahapatra and B. Ben Amor, “MedUnA: Language-guided Unsupervised Adaptation of Vision-Language Models for Medical Image Classification,” arXiv:2409.02729, 2024.
17. R. Jin, C.-Y. Huang, C. You, X. Li, “Backdoor Attack on Unpaired Medical Image-Text Foundation Models: A Pilot Study on MedCLIP,” arXiv:2401.01911, 2024.
18. Z. Sun, et al., “A scoping review on multimodal deep learning in medical imaging,” BMC Med. Imaging, vol. 23, no. 1, 2023.
19. W.-L. Wu et al., “Visual Entailment Based Contrastive Learning (VECL) for Medical Imaging,” MICCAI, 2025 (accepted).
20. R. Zheng, L. Yan, C. Gou, Z.-C. Zhang, J. J. Zhang and M. Hu, “Attention-based multi-modal fusion for radiograph and clinical note classification in COVID-19 diagnosis,” Information Fusion, vol. 75, pp. 168–185, 2021.
21. processing considering accuracy and performance. In Recent Advances in Computing Sciences (pp. 323-328). CRC Press.
22. Syedzagiriya S, “Efficient Computational Approaches for Solving Integro-Differential Equations in Engineering Applications”, Frontiers in Mathematical and Computational Research, pp. 26–34, May 2026, Accessed: Jul. 06, 2026.
23. G. Balaji, “Uncertainty-Guided Reduced-Order Models for Real-Time System Prediction”, Frontiers in Computational Science and Engineering , pp. 20–28, May 2026, Accessed: Jul. 06, 2026.
24. T. Z. Yang and L. Y. Yang, "Optimized Deep Learning Framework for Intelligent Data Analytics in Information Systems," Journal of Computer Applications and Information Technology, 2025.
25. S. Goundar and E. R. Kaburuan, "AI-Driven Cyber Risk Assessment: Protecting Against Cyberthreats Determined with Machine Learning," Journal of Wireless Networks and Communication Systems, 2025.