Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24887

# *The Research on Medical Image Processing Methods Based on Large Models*

**Chongwei Dong**

*College of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
*1120223542@bit.edu.com*

***Abstract.*** In recent years, with the widespread application of foundational large models in natural language processing and image re cognition, they are increasingly applied to medical imaging tasks. In medical scenarios, due to the inherently multimodal nature of the data, large models with joint image-text understanding capabilities demonstrate significant potential. This paper provides a comprehensive review of the typical applications of large models in the field of medical image processing, with a focus on the architectures, methodologies, and performance of representative multimodal medical large models. It compares their effectiveness in various medical imaging tasks. Additionally, it discusses the challenges faced in this field, including the difficulty of acquiring medical data, privacy protection during model training, and cross-modal information fusion. Finally, the paper outlines future directions for the development of large models in medical image processing. Recent studies suggest that medical foundation models may play a pivotal role in advancing intelligent imaging analysis systems.

***Keywords:*** Medicine, Large Language Models, Image Processing, Multimodal Models

## 1. Introduction

With the rapid advancement of artificial intelligence technologies, large models have demonstrated remarkable general capabilities in fields such as natural language processing and image recognition, particularly showing great potential in cross-modal processing and transfer learning. These models can efficiently process large-scale medical data. They can also be fine-tuned to adapt to domain-specific requirements in the medical field, such as lesion detection, disease diagnosis, and personalized treatment planning. As a result, they offer promising prospects for improving diagnostic efficiency and accuracy. Therefore, exploring large model applications in medical imaging is both practically and academically significant for overcoming the limitations of traditional technologies and advancing the intelligence level of healthcare services.

Currently, research on medical image processing based on large models has made notable progress. Multimodal vision-language models (VLMs) have emerged as a mainstream direction. However, existing studies still face numerous challenges: the scarcity of high-quality medical data, privacy requirements that restrict data sharing, hallucination phenomena in small-sample models, and potential performance bottlenecks caused by over-reliance on expert-annotated models. At present, medical large models are still in a stage of steady development.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24887

This paper provides a systematic review of the applications of large models in medical image processing, analyzing the architecture, methods, and performance of representative models, while also discussing existing challenges and envisioning future development directions. First, it examines the training frameworks and evaluation results of key models, comparing their performance across different tasks. Then, it explores practical challenges such as data scarcity, privacy protection, and model robustness. Finally, it outlines future development trends for large models in medical image processing, based on current technological advances. Through this comprehensive study, the goal is to provide both theoretical insights and methodological guidance to support the clinical implementation of large models in intelligent medical imaging analysis.

## 2. Research background and significance

### 2.1. Development of foundational pretrained models

Pretrained foundational models (PFMs), such as BERT and GPT-4, are artificial intelligence models that acquire general capabilities through large-scale data pretraining to learn universal feature representations [1]. The early development of foundational models was constrained by limited computational resources and insufficient dataset sizes, resulting in relatively small model scales and limited performance, primarily handling simple text-processing tasks. With the advancement of the internet, the rise of deep learning, and breakthroughs in natural language processing (NLP)—especially the introduction of the Transformer architecture—models like Google's BERT and OpenAI's GPT series have emerged, continuously improving in performance.

Subsequently, the exploration of multimodal technologies has enabled large models to process heterogeneous data such as text, images, and audio. For example, models like Flamingo and LLaVA utilize CLIP as an image encoder. Generative models capable of producing new, plausible outputs by learning statistical patterns from data have also developed rapidly, now applied to image generation, speech synthesis, and more.

Through large-scale pretraining, large models acquire strong representation learning capabilities. For cross-task and multi-task scenarios, they can be fine-tuned to quickly adapt to new tasks. As parameter and dataset sizes grow, large models also demonstrate scalable design features. Moreover, modular architectures support small-scale deployment needs. At present, large models exhibit excellent multimodal abilities, including multilingual translation, visual question answering, and image captioning.

### 2.2. Adaptation and fine-tuning of foundational models in medical scenarios

As general-purpose large models evolve, they have permeated various aspects of society, including the medical field. Compared to training general-purpose models from scratch, developing a medical large model is relatively less challenging. Most medical large models are built upon existing general-purpose model backbones and fine-tuned with domain-specific medical data, thereby enhancing their professional competencies in healthcare.

By leveraging prompts with cross-domain expressiveness, large language models can transfer knowledge across domains and improve generalization. They are also capable of recognizing new entities in few-shot or even zero-shot settings [2].

The healthcare field has increasingly recognized the limitations of relying on unimodal data. Medical data are inherently multimodal—clinical diagnoses often depend on multiple sources such as medical records, diagnostic reports (text), X-rays, ECGs, CT scans (images), and even doctor-

DOI: 10.54254/2755-2721/2025.PO24887

patient consultation recordings (audio). These information sources cannot be comprehensively captured by a single modality. Currently, large models are also evolving toward integrating natural language and vision, and existing multimodal models have shown theoretical feasibility in being fine-tuned and adapted for multimodal medical data.

### 2.3. Challenges and significance of medical large models

Medical datasets differ significantly from general-purpose datasets. Most existing datasets for medical large models focus on modalities like X-rays, CT, and MRI, while datasets for rare diseases remain scarce. Furthermore, medical data are highly sensitive, and limited data sharing affects model training and generalization. Additionally, medical data often require annotation by domain experts, which is costly and labor-intensive. In cases such as early disease prediction, it is difficult to obtain perfect labels.

Given these characteristics, medical large models face unique challenges. However, they can play a crucial role in medical image processing. Through intelligent analysis of medical images and cross-modal alignment (e.g., image-text), such models enable zero-shot reasoning and few-shot learning, thereby improving diagnostic capabilities for rare or complex conditions. They can also efficiently process massive volumes of imaging data and provide real-time feedback. The involvement of large models in medical image processing may reduce human errors and enhance diagnostic accuracy.

## 3. Analysis of representative techniques in medical image processing

### 3.1. Medical image processing techniques in medical large models

#### 3.1.1. Med3DVLM

Unlike 2D images that typically present a single cross-sectional view of a lesion, 3D images support volumetric measurement, spatial localization, and analysis of anatomical relationships. In clinical practice, many diagnostic tasks rely on 3D medical image processing, such as 3D reconstruction of lung CT scans, prenatal screening with 3D ultrasound, multi-sequence volumetric MRI data, and tumor monitoring through fusion of nuclear imaging with CT.

Med3DVLM is one of the few vision-language models specifically designed for 3D medical image analysis. This model is pretrained to handle volumetric data effectively. Unlike standard large models that primarily process 2D images, Med3DVLM utilizes a 3D image encoder called DCFormer, which decomposes traditional 3D convolutions into three parallel 1D convolutions along the depth, height, and width axes. This approach significantly reduces computational complexity from $O(k^3)$to $O(3k)$, enabling the model to process larger 3D images while preserving critical spatial details of volumetric medical data [3].

In vision-language pretraining, Med3DVLM adopts a sigmoid-based pairwise loss function instead of CLIP's softmax-based contrastive loss. This loss independently optimizes each image-text pair, avoiding global batch normalization. Such a design is more suitable for small-batch medical data and reduces semantic interference from similar samples. In both image and text retrieval tasks, the combination of DCFormer and the SigLIP approach significantly outperforms 2D-based models like PMC-CLIP. Moreover, under multi-sample settings, it achieves superior performance compared to M3D-LaMed, which relies on traditional softmax and global normalization.

DOI: 10.54254/2755-2721/2025.PO24887

Empirical evaluations show that Med3DVLM demonstrates superior performance in interpreting medical images with higher spatial precision. For instance, given the same image, Med3DVLM correctly identifies the lesion's location as the liver, whereas M3D-LaMed incorrectly responds with the pancreas. It also correctly recognizes pleural effusion in the right lung and detects abnormal masses in the stomach. In closed-ended visual question answering tasks—including imaging plane identification, organ recognition, and lesion characterization—Med3DVLM achieves the highest performance, with an average accuracy close to 80%. In contrast, M3D-LaMed averages around 76% and underperforms in all metrics compared to Med3DVLM, while RadFM only reaches baseline accuracy levels of around 20%.

Through efficient 3D feature encoding and robust cross-modal alignment, Med3DVLM significantly enhances the processing capabilities for 3D medical images. In open-ended visual question answering as well, Med3DVLM maintains its strong performance, achieving the best results across BLEU, ROUGE, METEOR, and BERTScore metrics compared to other models.

### 3.1.2. Med-Flamingo

To offer a more user-friendly and accessible model, LAION-AI released OpenFlamingo, an open-source Python-based model that inherits the strong multimodal capabilities of DeepMind's Flamingo. Med-Flamingo is trained based on OpenFlamingo-9B. In medical image processing, the model uses CLIP's ViT/L-14 vision encoder to convert X-rays, pathology slides, and other images into numerical visual embeddings. These embeddings capture key structures within the images to extract visual features relevant to medical analysis [4].

For dataset construction, Med-Flamingo introduces a unique interleaved image-text dataset (MTB), created using 4,721 medical textbooks and organized into 49 groups with Claude-1. This structure better aligns with how information is typically presented in the medical domain. Moreover, the model applies few-shot prompting to handle medical images and generate open-ended responses, rather than simply scoring options.

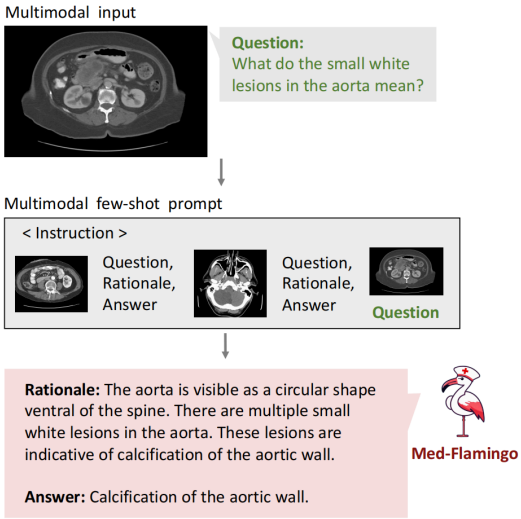


Figure 1: Illustrates the process of aortic wall calcification

In Figure 1, researchers ask the model about the white area in the image. Using few-shot prompts, Med-Flamingo divides its response into cause and conclusion: 'The aorta appears circular in front of

DOI: 10.54254/2755-2721/2025.PO24887

the spine. There are several small white lesions in the aorta. These indicate calcification of the aortic wall.' Thus, the model identifies the white area as a lesion in the aorta.

Across different datasets, Med-Flamingo demonstrates strong performance. On the VQA-RAD dataset, it achieves the highest scores in clinical evaluation, BERT-sim, and Exact-match metrics under few-shot conditions.

In contrast, on PathVQA (pathology-related images), all models score relatively low in clinical evaluation (scored directly by medical experts), highlighting a general weakness in current models for pathology image analysis. For cross-specialty tasks, where Exact-match becomes ineffective, evaluation relies heavily on expert scoring, and Med-Flamingo still performs well in the few-shot setting.

### 3.1.3. VILA-M3

Most publicly available medical datasets used for model training and evaluation are static and not tailored to the training requirements of large vision-language models (VLMs). General-purpose models like GPT-4 rely heavily on internet-derived knowledge and often lack the specialized expertise required for medical contexts. Existing medical VLMs such as Med-Gemini also lack expert model integration and cannot fully meet the demands of modern deep learning in healthcare. To address this, VILA-M3 was developed with the goal of incorporating expert-level medical knowledge into VLMs.

The training pipeline begins with standard vision pretraining using OpenAI's CLIP-L model as the vision encoder. The VILA framework employs an autoregressive architecture for vision-language pretraining. Fine-tuning is then conducted on a hybrid dataset. Finally, domain-specific expert models are integrated into the decision-making and response-generation process. Depending on the input image and query, VILA-M3 dynamically invokes appropriate expert modules, such as VISTA3D for volumetric segmentation in CT, MONAI BRATS for MRI brain tumor segmentation, and TorchXRayVision for chest X-ray classification [5]. This integration enables VILA-M3 to handle diverse medical tasks, ranging from segmentation to diagnostic classification.

Anomalous structures in medical images often exhibit fine-grained visual characteristics (e.g., edge morphology, tissue contrast), which basic large models like GPT-4o struggle to detect with image-text training alone. VILA-M3, by incorporating expert models capable of pixel-level annotations, generates more accurate and detailed responses. When no expert model is used, both VILA-M3 and GPT-4o fail to identify small tumors in certain images. After integrating expert modules, both models successfully locate the tumor and highlight it in red.

A comparison among different VILA-M3 parameter versions (from 3B to 40B), GPT-4o, Med-Gemini, and TorchXRayVision was performed using F1 scores. While VILA-M3 exhibits some fluctuation across parameter scales, it maintains an average F1 score close to 50. With expert model integration, VILA-M3 sees a 3–5% improvement in overall F1 scores, indicating enhanced diagnostic accuracy. VILA-M3 also shows notable advantages over general models such as GPT and Gemini, outperforming them by 2–10% on tasks involving fracture detection, pneumothorax, and lung opacity. Compared to the expert model TorchXRayVision, VILA-M3 achieves a slightly higher average F1 score, demonstrating its strong potential in medical image processing.

### 3.1.4. XrayGPT

XrayGPT is developed based on two datasets: MIMIC-CXR (general clinical radiology) and OpenI (focused on radiology reports). It employs a frozen medical visual encoder, MedCLIP, and performs

DOI: 10.54254/2755-2721/2025.PO24887

feature alignment between vision and language representations via a linear transformation layer to achieve effective multimodal fusion [6].

Visual Feature Encoding Formula:The encoder Eimg extracts low-level features from an X-ray image x. These high-dimensional visual features are then projected into a 512-dimensional embedding space through a linear projection head fv, facilitating downstream processing and converting raw images into numerical representations.

Cross-modal Alignment Formula:The output feature vector from the visual encoder undergoes a trainable linear transformation to produce aligned visual-language embeddings, ensuring semantic consistency across modalities. Although XrayGPT excels at radiographic summarization, its performance may degrade when applied to modalities beyond X-rays.

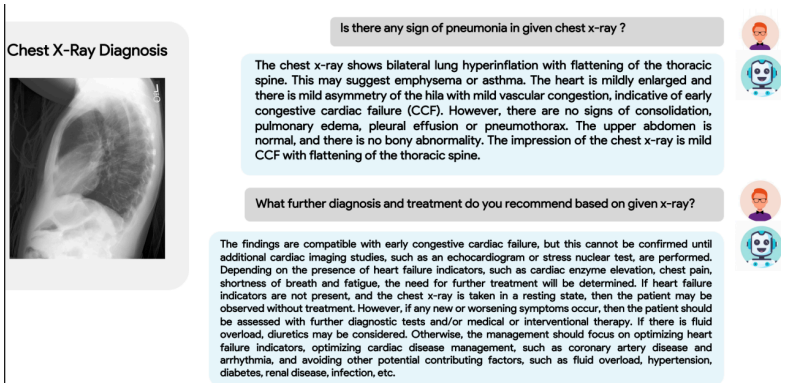


Figure 2: XrayGPT's summary and impression of a chest X-ray

In Figure 2, the model is prompted to summarize a given chest X-ray image. XrayGPT accurately rules out three critical pathological conditions—focal consolidation, pleural effusion, and pneumothorax. The final impression generated by the model reads: "No acute cardiopulmonary abnormality identified."

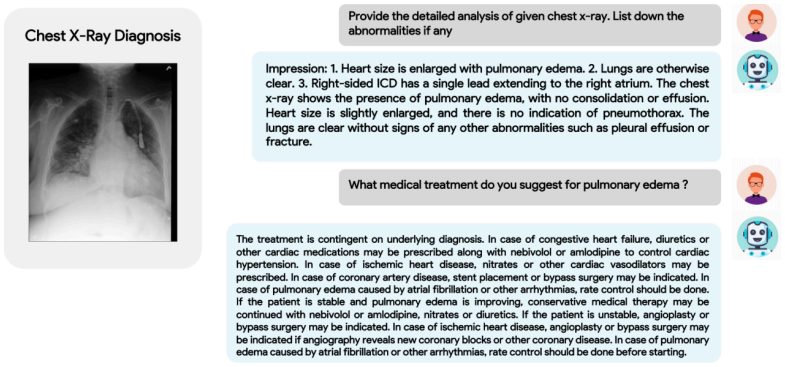


Figure 3: Disease identification by XrayGPT

As shown in Figure 3, XrayGPT accurately identifies cardiomegaly accompanied by pulmonary edema, a cardiovascular-related condition. Furthermore, the model suggests diverse treatment strategies surrounding potential etiologies, including pharmacotherapy, interventional procedures, and surgical interventions, showcasing its potential as a clinical decision support tool.

DOI: 10.54254/2755-2721/2025.PO24887

## 3.2. Comparative analysis of representative techniques

Table 1: Comparison of multimodal models in medical image processing

| Model | Evaluation | Dataset | Technique | Limitations |
|---|---|---|---|---|
| Med-Flamingo | Generates open-ended answers; evaluated by three physicians, BERT similarity, and exact match | Unique interleaved image-text dataset (MTB), created from a set of medical textbooks using Claude-1 | Built upon OpenFlamingo-9B; leverages multimodal in-context learning; follows the original Flamingo model; trained by minimizing a joint objective | Unstable generation quality; presence of hallucinations |
| MEDIFICS | Binary response matching (score 0 or 1 based on full or partial match) | Simulated medical dialogues based on metadata from datasets like ISIC, ROCO, MURA | Fine-tuned IDEFICS-9B-instruct using QLoRA; dynamically selects and applies task-specific expert models [7] | Limited robustness; insufficient adaptability |
| MedVLM-R1 | Score-based comparison among models like Qwen2-VL, HuatuoGPT-Vision on in-domain (MRI) and out-of-domain (CT, X-ray) tasks | Evaluated on HuatuoGPT-Vision benchmark covering VQA-RAD, SLAKE, PathVQA, OmniMedVQA, and PMC-VQA | Uses Group Relative Policy Optimization (GRPO) instead of standard supervised fine-tuning; aims to improve natural language reasoning and output transparency [8] | Poor convergence on non-MRI modalities; over-reliance on closed-set VQA; prone to shallow or hallucinatory reasoning |
| VILA-M3 | Evaluated on RadVQA, SLAKE, and PathVQA using BLEU-4, ROUGE, GREEN, F1, and comparison with Med-Gemini and task-specific SOTA baselines | RadVQA, SLAKE, and PathVQA for visual question answering in medical imaging | Integrates expert models like TorchXRayVision for X-ray classification, MONAI BRATS for MRI tumor segmentation, VISTA3D for CT volumetric segmentation | Prone to overfitting; dependent on the availability and accuracy of expert models |
| UniMed-CLIP | Evaluated on multiple datasets (MM-Retinal, MIMIC-CXR, PMC-OA, ROCOv2) and compared with general-purpose models | Leverages LLMs to convert high-quality labeled data into 5.3 million image-text pairs across six modalities | Zero-shot evaluation; employs multi-captioning strategy—randomly selects one caption per image to increase training diversity [9] | — |
| Med3DVLM | Assessed via Recall@K (R@K), BLEU, ROUGE, METEOR, BERTScore | Public M3D dataset | Uses DCFormer to reduce computational complexity while preserving spatial details; SigLIP enhances contrastive learning; MLP-Mixer fuses low- and high-level visual features | Despite performance gains in open-ended VQA, still shows inaccuracies in lesion localization and answer precision |
| XrayGPT | Progressive evaluation using an expanding set of datasets | MIMIC-CXR (377,110 images, 227,827 reports) and OpenI (6,459 images, 3,955 reports) with high-quality, interactively summarized data | Uses MedCLIP as the vision encoder, built upon the Vicuna architecture | Optimized for X-ray analysis, but limited to this modality |

All models in Table 1 are multimodal, capable of processing both images and text. In terms of model origin, they can be broadly categorized into fine-tuned general-purpose models and those pretrained on medical-specific datasets. Regarding task specificity, some models are designed for specialized tasks—for example, XrayGPT focuses on chest X-rays, while Med3DVLM is optimized for 3D medical imaging. Others demonstrate strong generalization across diverse datasets.

Overall, multimodal modeling has become a central trend in the development of medical foundation models. Medical image processing is increasingly considered an essential capability of such models. Whether fine-tuned from existing general-purpose models or pretrained from scratch using medical datasets, these models show enhanced performance in handling complex medical imaging tasks.

## 4. Challenges and limitations

Despite the remarkable progress in medical image processing driven by large-scale medical foundation models, several critical challenges and limitations hinder their practical application and large-scale deployment. These challenges are summarized as follows:

DOI: 10.54254/2755-2721/2025.PO24887

### 4.1. Data scarcity and privacy concerns

The acquisition of high-quality medical data has long been a bottleneck, particularly for rare disease cases, which significantly limits the diversity and robustness of model training. Some models, such as Med-Flamingo and UniMed-CLIP, attempt to mitigate this issue by constructing proprietary databases. Moreover, ensuring patient data privacy remains a critical concern in model development and deployment. Due to the sensitive nature of medical data, model training must comply with relevant regulations. In recent years, federated learning and synthetic data generation have gained attention as potential solutions to enable data sharing and collaborative training while preserving privacy.

### 4.2. Limited generalization and hallucination issues

Nevertheless, despite their success, current models remain prone to hallucinations—producing seemingly plausible but factually incorrect outputs—especially under few-shot or open-domain scenarios. For example, MedVLM-R1 incorrectly localized a tumor to the right side in a test case, contradicting the actual diagnosis. Med3DVLM has also exhibited significant misinterpretations in certain multimodal inputs. Such errors could lead to severe consequences in clinical settings. Therefore, improving model robustness and reliability remains a crucial research focus.

### 4.3. Dependence on expert models

Models such as MEDIFICS and VILA-M3 rely heavily on domain-specific expert systems to enhance their performance. While this strategy allows rapid knowledge transfer and improves initial model capabilities, it can also limit the model's upper performance bound, as it becomes constrained by the capabilities of the expert models themselves.

### 4.4. Lack of deep medical understanding

Most current models remain at the surface level in understanding medical texts or images. They often lack the ability to model causal relationships between complex diseases, therapeutic pathways, and long-term treatment outcomes. In cases involving comorbidities, models struggle to integrate multiple factors, resulting in diagnostic or therapeutic suggestions that lack systematic reasoning, specificity, and interpretability. When faced with tasks such as medication selection, efficacy evaluation, and treatment duration planning, models still fall short and require more domain-specific training and data.

### 4.5. Barriers to clinical deployment

Many medical foundation models are still in the experimental or early validation stages. Seamless integration with hospital information systems, electronic health records (EHRs), and other clinical workflows is far from mature. Clinical deployment faces numerous obstacles, including rigorous validation protocols, building physician trust, regulatory compliance, and accountability mechanisms—all of which are necessary for real-world adoption but remain underdeveloped.

DOI: 10.54254/2755-2721/2025.PO24887

## 5. Research trends and future directions

In response to the current challenges, both academia and industry are actively exploring multiple improvement strategies to advance the practicality and intelligence of medical foundation models.

Model interpretability is a key factor influencing clinical adoption. Incorporating mechanisms such as attention visualization, case-based reasoning, and causal graphs can enhance the traceability and transparency of model decisions. Additionally, incremental learning is gaining attention as a means to allow models to continually assimilate new knowledge while avoiding catastrophic forgetting, thereby maintaining long-term adaptability.

In the foreseeable future, medical foundation models are anticipated to converge towards unified multimodal architectures that integrate heterogeneous data sources such as text, medical images, and speech. These models aim to emulate the multi-source decision-making process of clinicians, leading to more accurate diagnoses and comprehensive treatment assistance.

Integrating structured medical knowledge—such as ICD-10, SNOMED CT, and UMLS—with large language models offers a promising direction to enhance domain specificity and factual consistency. Current research explores strategies including knowledge graph injection, medical term annotation, and graph neural networks combined with Transformer architectures to deeply fuse language and domain knowledge.

To mitigate hallucination problems, Retrieval-Augmented Generation (RAG) has emerged as a widely studied approach. RAG enhances response reliability by retrieving relevant literature or knowledge from external databases prior to response generation, ensuring evidence-based and up-to-date answers. This architecture is particularly well-suited to the highly dynamic knowledge landscape of the medical domain, and holds significant promise for future applications [10].

## 6. Research applications

Language models applied to medical image processing have been widely adopted in various domains, including fundamental disease-assisted diagnosis, medical image analysis and processing, and medical image synthesis. Moreover, large foundation models for medical image processing exhibit strong potential for cross-disciplinary applications.

In the field of health management, These models enable longitudinal monitoring and predictive analysis of patients' medical imaging data, facilitating early risk assessment and personalized interventions, providing valuable support for insurance claims and personalized health management.

In medical device development, large-scale medical image processing models assist in optimizing imaging equipment. By analyzing vast amounts of medical imaging data, they help improve imaging algorithms, enhancing resolution, clarity, and accuracy. This enables medical imaging devices to play a greater role in early disease screening and detection of small lesions.

During drug development, medical image processing techniques based on large models facilitate the analysis of drug mechanisms and effects within the body. Leveraging medical imaging data, these models enable observation of drug impact on pathological tissues and assist in evaluating drug efficacy and safety.

## 7. Conclusion

With the continuous advancement of artificial intelligence technologies, multimodal foundation models have demonstrated broad application prospects in the field of medical image processing. This paper systematically reviews current representative large-scale medical image processing

DOI: 10.54254/2755-2721/2025.PO24887

models, including Med3DVLM, Med-Flamingo, VILA-M3, and XrayGPT, analyzing their architecture design, data preprocessing, task adaptation capabilities, and cross-modal alignment mechanisms in depth. Studies show that these models demonstrate marked advancements in vision-language alignment, open-ended visual question answering, and the incorporation of domain-specific medical knowledge. Some models have in some cases outperformed traditional methods in complex tasks such as 3D image understanding, lesion localization, and cross-modal reasoning.

At the same time, this paper highlights multiple challenges faced by large models in medical image processing, including scarcity of medical data, data privacy protection, hallucination generation, difficulties in multimodal semantic alignment, and lack of interpretability in model inference. Nonetheless, these issues can be partly mitigated through the introduction of domain expert models, multimodal fine-tuning techniques, structural optimization, and cross-modal pretraining strategies.

Future development of large models for medical image processing should focus on protecting patient privacy, enhancing model generalization, improving diagnostic accuracy and interpretability, and gradually achieving wide clinical deployment. Moreover, integrating cutting-edge approaches such as federated learning, generative adversarial networks (GANs), and medical knowledge graphs will further promote the intelligent, personalized, and efficient evolution of intelligent medical imaging analysis systems.

## References

[1] Ce Zhou, Guangjing Wang and Qiben Yan. 2023. Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv preprint arXiv: 2302.09419

[2] Ziyuan Qin, Huahui Yi, Qicheng Lao, Kang Li. 2023. Medical Image Understanding with Pretrained Vision Language Models: A Comprehensive Study. arXiv preprint arXiv: 2209.15517

[3] Yu Xin, Gorkem Can Ates, Kuang Gong, Wei Shao. 2025. Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis. arXiv preprint arXiv: 2503.20047

[4] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P. & amp; Rajpurkar, P.. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. https: //proceedings.mlr.press/v225/moor23a.html

[5] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Greg Heinrich, Yee Man Law, Benjamin Simon, Stephanie Harmon, Stephen Aylward, Marc Edgar, Michael Zephyr, Song Han, Pavlo Molchanov, Baris Turkbey, Holger Roth, Daguang Xu. 2025. VILA-M3: Enhancing Vision-Language Models with Medical Expert Knowledge. arXiv preprint arXiv: 2411.12915

[6] Omkar Thawakar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, Fahad Shahbaz Khan. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. arXiv preprint arXiv: 2306.07971

[7] E. -T. Said, A. E. A. Soufiane and E. -T. Jamal, "MEDIFICS: Model Calling Enhanced VLM for Medical VQA, " 2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 2024, pp. 1-6, doi: 10.1109/ICDS62089.2024.10756285. https: //ieeexplore.ieee.org/abstract/document/10756285

[8] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, Daniel Rueckert. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv preprint arXiv: 2502.19634

[9] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan. 2024. UniMed-CLIP: Towards a Unified Image-Text Pretraining Paradigm for Diverse Medical Imaging Modalities. arXiv preprint arXiv: 2412.10372

[10] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, Huaxiu Yao. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. arXiv preprint arXiv: 2410.13085