·专家述评·

# 医学影像大模型的演进、技术架构与临床展望评述*

李璐[1,2] 孙怀强[1,2]

(四川大学华西医院 1. 放射科;2. 放射影像研究所,四川 成都 610041)

【摘要】 近年来,人工智能在医学影像分析领域正经历从“专用模型”向“基础模型”范式的转变。传统单任务模型高度依赖专家标注且缺乏跨任务泛化能力,而医学影像大模型(LMIMs)通过海量多模态数据自监督预训练,仅需少量微调即可适应多种下游任务,是迈向医疗通用人工智能的关键路径。本文系统评述了医学影像大模型的最新研究进展。首先,将现有模型分为视觉基础模型、视觉-语言大模型以及通用与智能体模型三大类。其次,深入剖析了核心架构(如大核卷积神经网络、Vision Transformer 及其混合架构)、对比学习、掩码建模等预训练学习范式。最后,探讨了数据构建与跨中心泛化的落地挑战,重点梳理了其在肿瘤等重大疾病中的临床应用潜力,并对结合因果推理、检索增强生成等技术破解部署瓶颈进行了展望。综上,医学影像大模型代表了医学人工智能发展的重要里程碑,未来有望深刻变革诊断流程,提升诊疗质量与效率,最终惠及全球患者健康。

【关键词】 医学影像大模型;基础模型;自监督学习;视觉-语言模型;通用人工智能

【中图分类号】 R445 【文献标志码】 A **DOI**:10. 3969/j. issn. 1672-3511. 2026. 04. 001

# Foundation models in medical imaging: a review of evolution, architecture and clinical prospects

**LI Lu[1,2], SUN Huaiqiang[1,2]**

(1. *Department of Radiology, West China Hospital, Sichuan University, Chengdu* 610041, *China*;
2. *MR Research Center, West China Hospital, Sichuan University, Chengdu* 610041, *China*)

**【Abstract】** In recent years, artificial intelligence in medical image analysis has been experiencing a paradigm shift from "task-specific models" to "foundation models". Traditional single-task models rely heavily on expert annotations and lack cross-task generalization capabilities. In contrast, Large Medical Imaging Models (LMIMs), pre-trained on massive multimodal data via self-supervised learning, can be adapted to a wide range of downstream tasks with only minimal fine-tuning, representing a critical pathway toward artificial general intelligence in healthcare. This article systematically reviews the latest research progress on LMIMs. First, existing models are categorized into three main classes: vision foundation models, vision-language large models, and generalist and agent models. Second, we provide an in-depth analysis of the core architectures (such as large-kernel Convolutional Neural Networks, Vision Transformers, and their hybrid architectures) and pre-training learning paradigms, such as contrastive learning and masked modeling. Finally, we discuss the practical challenges of data construction and cross-center generalization, highlight their clinical application potential in major diseases such as oncology, and provide perspectives on overcoming deployment bottlenecks by integrating technologies like causal inference and retrieval-augmented generation. In summary, LMIMs represent a significant milestone in the

**基金项目**:四川省自然科学基金项目(2024NSFSC0656)

**执行编委简介**:孙怀强,四川大学华西医院放射影像研究所副研究员,博士研究生导师,磁共振技术研究室主任。入选四川省科学与技术带头人后备人选,四川省卫健委科学与技术带头人后备人选,“天府青城计划”青年科技人才,四川大学“双百”人才工程(B 类),中国科协青年人才托举工程。以第一或通信作者在 *JAMA Psychiatry*,*Radiology*,*Ebiomedicine* 等高水平期刊发表 SCI 论文 30 篇。先后主持国家自然科学基金青年、面上项目,科技部重点研发专项课题,四川省自然科学基金面上项目,四川大学青年领军人才培育项目,华西医院 1·3·5高水平发展项目。研究成果获中华医学科技一等奖,华夏医学科技奖一等奖,四川省自然科学一等奖。主要研究方向为磁共振神经影像以及医学影像人工智能。E-mail:sunhuaiqiang@scu. edu. cn

**引用本文**:李璐,孙怀强. 医学影像大模型的演进、技术架构与临床展望评述[J]. 西部医学,2026,38(4):469-477. DOI:10. 3969/j. issn. 1672-3511. 2026. 04. 001

development of medical artificial intelligence, holding the promise of profoundly transforming diagnostic workflows, improving the quality and efficiency of clinical care, and ultimately benefiting global patient health.

【Key words】 Large medical imaging models; Foundation models; Self-supervised learning; Vision-language models; Artificial general intelligence

过去十年,以卷积神经网络(Convolutional neural networks, CNNs)为代表的深度学习技术在医学影像分析领域取得了显著成就[1]。在肺结节检测、视网膜病变筛查等特定任务中,这类模型的诊断准确率已达到甚至超越人类专家水平[2]。然而,当前医学影像人工智能仍主要处于“专用人工智能”阶段,面临着诸多制约因素。首先,单任务模型高度依赖大规模、高质量的专家标注数据,在数据获取与标注成本方面存在显著瓶颈;其次,模型往往缺乏跨中心、跨设备的泛化能力。这些局限性使得传统深度学习模型难以有效融入复杂多变的临床工作流程。近年来,随着 Transformer 架构[3]的引入以及大语言模型(Large language models, LLMs)如 GPT-4[4]的突破性进展,人工智能正在经历从“监督学习”向“自监督预训练”的范式转移。受 LLMs 在大规模无标注文本上习得通用语言能力的启发,视觉领域迅速涌现出“基础模型”[5]的概念。与传统模型不同,基础模型通过在海量多样化数据上进行预训练,学习数据的通用特征表示,仅需少量微调甚至零样本提示[6-7],即可适应多种下游任务。

“基础模型”概念最早由斯坦福大学以人为本人工智能研究院提出,其核心定义为:“通过自监督或半监督学习方式在大规模数据集上完成训练,可适配各类下游任务的基准模型”[5]。该范式以深度神经网络和自监督学习为理论基础,通过扩展数据规模与模型参数实现跨领域广泛应用。当这一理念应用于医学领域时,便形成了医学影像大模型(Large medical imaging models, LMIMs)。LMIMs 特指在大规模、多模态医学数据上通过自监督预训练,能够学习人体解剖结构与病理特征的通用表征,并具备跨任务迁移能力与涌现能力的新一代医学影像分析模型。当前医学影像大模型研究主要集中在两个方向:一是视觉基础模型,如医学版的 SAM[8-9],旨在解决通用的解剖结构识别与分割问题;二是视觉-语言大模型,通过对齐影像与放射学报告,使模型能够理解复杂临床语境,具备生成诊断报告甚至进行交互式视觉问答的能力[10-11]。这种通用的感知与认知能力不仅弥合了视觉信息与临床语义之间的鸿沟,更打破了传统人工智能“一个任务一个模型”的局限,被认为是迈向医疗通用人工智能的关键一步[12]。本文系统评述了医学影像人工智能从“专用模型”向“基础模型”转型的最新技术进展与临床应用,可以预见的是,随着模型能力的持续进化与多模态融合的深化,医学影像大模型不仅能精准识别病灶、量化病变进展,更能理解病历语境、辅助临床决策,最终推动医学影像从“辅助诊断”迈向“协同诊疗”,为医疗通用人工智能的实现奠定坚实基础。

## 1 模型类型及适用的下游任务

医学影像大模型正经历从单一视觉向多模态及智能体的演进,其主要分为视觉基础模型、视觉-语言大模型以及通用与智能体模型三大类。

1.1 视觉基础模型 此类模型专注像素级信息,通过自监督学习提取通用特征,微调后即可适配分割、检测等多种下游任务。架构演进上相比传统方法,视觉 Transformer(Vision Transformer, ViT)展现出强大优势。BrainSegFounder[13]证明了其在 3D 神经影像分割中的优势;ScarNet[14]在心肌瘢痕量化任务上,展现了超越传统方法的精度;DinoV2[15]和 RadDino[16]在少样本分类中表现优异;LVM-Med[17]通过百万级数据预训练,展现了跨域多器官筛查的强大特征表达能力。在分割任务上 SAM[8]确立了新通用范式。MedSAM[9]和 MedSAM2[18]实现了全身器官零样本分割;SAM2-3dMed[19]突破时空维度,有效解决 3D 连续层面的器官分割与血管追踪难题;而最新的 Iris 模型[20]引入“上下文学习”,仅凭少量参考图像即可实时分割罕见病灶,挑战了传统的“预训练-微调”范式。

1.2 视觉-语言大模型 视觉-语言大模型(Vision-language models, VLMs)旨在弥合像素特征与文本语义的鸿沟,分为判别式和生成式两类。判别式模型侧重图文对齐,主要用于图像检索、零样本分类及异常检测。例如受 CLIP[21]启发,MedCLIP[22]和 Biomed-CLIP[23]建立了通用医学语义空间,常用于跨模态医学图像检索;CXR-CLIP[24]和 Mammo-CLIP[25]针对胸部 X 线肺炎筛查和乳腺癌分类进行了特异性优化;CT-CLIP[26]和 EchoCLIP[27]则实现在无标注数据下,通过自然语言即可定位 3D 肺结节或检测心脏异常。生成式模型则集成大语言模型的生成能力,能够执行报告生成,视觉问答等任务。RadFM[11]在 X 线报告

自动撰写上优于 GPT-4V；Med-Gemini[28]能处理复杂的视觉问答（如“解释该 CT 扫描中阴影的病理意义”）；Med-PaLM M[29]进一步树立通用标杆，证明了单模型可同时处理皮损识别、报告生成至基因组学分析等 14 种截然不同的任务。最新研究正在突破以往 2D 或静态分析的瓶颈。M3D-LaMed[30]提出高效 3D 空间池化感知器，实现了原生对齐 3D 视觉特征与文本语义的直接对齐；MERLIN[31]则结合 CT 与电子病历进行预训练，不仅能理解当前的影像表型，还能结合患者长期的病史轨迹进行跨模态推理，在 5 年慢性病预测等预后任务上展现了显著的临床价值。

1.3 通用与智能体模型 研究焦点正转向能处理跨领域异构数据的通用模型与具备自主规划能力的智能体。模型利用统一架构处理多模态数据，如 Perceiver IO[32]和 VATT[33]结合超声与心音进行综合诊断，CoMET[34]则从影像和临床记录中预测生命周期健康轨迹。TGSAM-2[35]实现了自然语言指令引导的动态超声实时目标追踪。前沿领域受“视觉-语言-动作”模型启发，其将感知与动作预测结合以辅助手术机器人，推动医学大模型从“数字世界”的诊断迈向“物理世界”的诊疗干预。

## 2 医学影像大模型的技术框架

2.1 核心架构

2.1.1 视觉基础模型常见架构 视觉基础模型的架构设计呈现多元并进态势。CNN 通过归纳偏置在医学影像应用中达到当前最优水平[36]，而 ViT 引入的全局建模范式也展现出强大的竞争力[37-39]。CNN 擅长通过空间卷积提取局部特征，但浅层网络难以理解全局上下文；相反，ViT 通过分块自注意力机制学习长距离依赖，但易丢失局部空间信息，且在识别微小病灶方面存在局限性[40]。因此，通过混合架构同时捕捉全局与局部内容成为重要研究方向[41]。当前主流架构主要分为三大类：①卷积神经网络（CNNs）：传统 CNN 受限于小卷积核，但“大核卷积”在保留平移不变性等归纳偏置的同时，大幅扩大了感受野，能够像 Transformer 一样捕捉长距离依赖。在处理高分辨率 3D 医学影像时，大核 CNN 通常比 ViT 具有更高的推理效率和更低的显存占用。例如，STU-Net[42]证明了纯 CNN 架构扩展至十亿级参数时，依然遵循缩放定律实现性能持续增长。②Vision Transformer 架构：为克服 CNN 在长距离依赖上的局限，ViT 将图像切分为展平的补丁作为输入令牌[43]。利用多头自注意力机制（MSA），模型在第一层即可捕捉全局上下文[44-45]。多个基于 ViT 的系统在分类分割等应用中表现良好[38, 46-49]，但通常需要比 CNN 更多的数据量进行预训练才能收敛。③混合视觉架构（HVTs）：鉴于医学数据的稀缺性与高维特性，将 CNN 与 ViT 结合的混合架构是当前的主流趋势[50]。根据结合方式，其拓扑结构主要分为串行结构[51-52]、并行结构[53]和分层结构。

2.1.2 视觉-语言大模型常见架构 视觉-语言模型（VLMs）通过跨模态交互打破了单一视觉感知的局限，旨在处理和推理视觉与文本两种模态。这些模型支持图像描述生成、跨模态检索、视觉问答和图像生成等多项任务，现有架构主要分为三类：①基于编码器的跨模态对齐（Encoder-based Cross-modal Alignment）：此类模型在统一特征空间中实现图像与文本的表征对齐，通过对比损失或图文匹配损失拉近匹配数据的特征距离。除开创性的 CLIP[21]及进一步优化的 ALIGN[54]、CLOOB[55]和 DeCLIP[56]外，医学领域的 MedCLIP[22]、PubMedCLIP[57]、ConVIRT[58]和 MGCA[59]等模型在基于文本的医学影像检索及零样本辅助诊断任务中展现出显著优势。②基于编码器的多模态注意力（Encoder-based Multimodal Attention/Fusion）：此类架构在统一编码器中整合视觉和文本输入，利用自注意力机制在编码器内直接建模跨模态交互，捕捉上下文关系。例如，VisualBERT9[60]将图像块和文本标记输入共享编码器，在视觉问答等任务中表现出色。在医学领域，这种深度融合方法（如 M3AE[61]）对于处理临床影像与病历记录之间的细微关联至关重要。③基于编码器-解码器的多模态整合（Encoder-decoder/generative models）：此类架构旨在实现条件生成，能够基于多模态输入产生自然语言描述或合成图像，广泛应用于医学报告生成等任务，其通常采用“视觉编码器+桥接模块+大语言模型解码器”的结构，通过桥接模块将视觉特征转化为 LLM 可理解的 token。此类模型除 SimVLM[62]、VisualGPT[63]和 Flamingo[64]外，在医学领域，LLaVA-Med[65]和 RadFM[66]等模型沿用此架构，经微调后实现了高质量的医学报告自动生成和多轮问诊对话。

2.2 学习范式 医学影像分析长期面临着“数据缺乏”与“标注稀缺”的双重挑战[67]，通过医学基础模型的学习范式革命性改变，极大缓解了这一瓶颈。

2.2.1 自监督预训练 自监督学习（Self-supervised learning, SSL）旨在通过设计辅助任务，从海量无标签数据中学习挖掘图像的内在结构特征。当前医学

影像领域的 SSL 范式主要分为对比学习和掩码建模两大类。对比学习的核心思想是通过构建正负样本对，在同一空间中拉近相似样本，推远不相关样本。受 SimCLR[68] 等启发，Azizi 等[69] 提出的多视图对比学习方法，利用电子病历中的元数据作为弱监督信号来构建正样本对，证明了 SSL 预训练能显著提升模型在皮肤病与胸部 X 线诊断中的鲁棒性。掩码图像建模（Masked image modeling，MIM）受 BERT[70] 的启发，通过随机掩盖图像部分区域（如 75%），迫使模型基于可见部分重建缺失的像素或特征（图 1A）。这种生成式的自监督任务迫使模型深度理解全局解剖结构与上下文关系，而非仅仅关注局部纹理[71-72]。

2.2.2　视觉-语言对齐目标　为跨越像素和临床语义的鸿沟，实现视觉特征与文本嵌入的对齐，多模态大模型通常通过多种预训练目标的组合来实现[73]。图文对比学习（Image-text contrastive learning，ITC）旨在建立全局语义对齐。通过在共享潜在空间最大化成对“影像-报告”的余弦相似度并推离非匹配对（图 1B），MedCLIP[22] 和 BiomedCLIP[23] 等模型凭借 ITC 任务，在零样本检索与分类中展现卓越的性能。图文匹配（Image-text matching，ITM）为了弥补对比学习在细粒度特征交互上的不足，模型需要判断输入的一对图像和文本是否真正匹配，使其能捕捉更细微的局部特征对应关系，例如判断报告中的“右上肺实变”是否确实出现在影像的对应区域[74]。掩码语言建模（Masked language modeling，MLM）在生成式架构中，MLM 训练模型以视觉特征为条件预测文本序列的下一个 token，这是赋予 RadFM[11] 等模型“生成诊断报告能力”的核心机制。

## 3　医学影像大模型在临床落地中的挑战与前沿展望

尽管医学影像大模型在底层架构与技术指标上取得了显著进展，但要重塑现有的临床工作流，仍需跨越多重现实壁垒。未来的研究重心必须从单一算法指标的刷榜，全面转向增强临床可用性、加强专病应用深度以及通用人工智能的构建。

3.1　高质量数据基石与跨中心泛化难题　高质量、大规模且具备长程随访记录的数据集是大模型迈向临床应用的前提。当前，基础模型的预训练高度依赖三类代表性数据集：①单模态影像数据集。RadImageNet[75] 包含超过 100 万张放射学图像，涵盖多种成像模态（CT、MRI、X 线），为视觉预训练提供了丰富的解剖结构信息。CheXpert[76] 和 ChestX-ray8[77] 分别提供 22 万和 11 万张胸部 X 线片，支持多标签肺部疾病分类任务。EMBED[78] 和 OPTIMAM[79] 则专注于乳腺癌筛查，前者包含 340 万张乳腺 X 线影像，具有突出的种族多样性。RSNA 系列挑战数据集（如脑出血[80]、肺栓塞[81]、颈椎骨折[82]）提供了专家标注的病灶定位信息，推动了检测算法的发展。SA-Med2D-20M[83]、TotalSegmentator[84] 和 AbdomenCT-1K[85] 为分割任务提供了数百万级的像素标注，覆盖全身多器官。②多模态图文对数据集。MIMIC-CXR[86] 是目前最大的公开图文对数据集，包含 37 万张胸部 X 线及其对应的放射学报告，为视觉-语言模型预训练提供了核心资源。CANDID-PTX[87] 进一步补充了气胸病例的细粒度标注。PMC-OA 和 PMC-CaseReport 从 PubMed Central 医学文献中提取了数百万对图文数据，覆盖了多种成像模态和临床场景，为构建通用医学多模态模型提供了丰富的知识来源。③视觉问答数据集。VQA-RAD[88] 和 SLAKE[89] 为医学视觉问答任务提供了基准，包含数千个影像-问题-答案三元组，要求模型不仅理解影像，还需进行临床推理。这些数据集推动了生成式 VLMs 向交互式临床决策支持系统的演进。

尽管现有的公开数据集为大模型的早期开发提

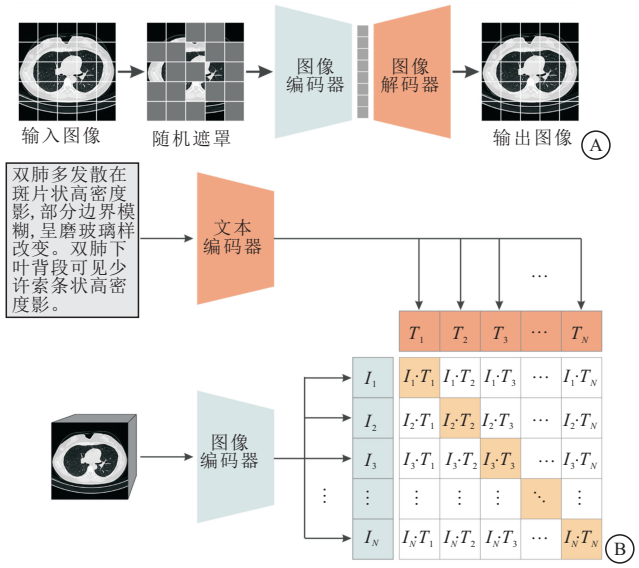


**图 1　医学影像大模型预训练学习范式**

**Figure 1　Large medical imaging model pre-training paradigm**

注：A. 掩码图像建模：输入的图像经过随机掩码处理，图像编码器仅对可见的图像块进行特征提取，随后图像解码器利用这些特征重建原始图像。此阶段旨在通过图像自身的上下文信息学习视觉特征表示。B. 医学影像-文本对比学习：展示了双塔结构的跨模态对齐过程。文本编码器提取放射科报告的语义特征（T1-TN），图像编码器提取图像的视觉特征（I1-IN）。右侧矩阵展示了图像嵌入与文本嵌入之间的相似度计算，旨在拉近成对的图像与文本在特征空间中的距离。

供了重要支撑，但要推动其在复杂临床场景中的广泛落地，当前的数据集无论在规模、维度还是真实世界代表性上仍显不足。未来医学影像人工智能的纵深发展，迫切需要打破数据孤岛，跨机构联合构建更多源自真实临床工作流、包含详尽临床金标准、具有长程预后随访记录以及融合多组学信息的高质量大队列数据集。此外，临床落地的核心挑战还在于“数据偏见”与“分布偏移”[90]。当前数据来源的单一性（主要来自北美和欧洲医疗机构）导致模型在其他人群中表现下降。医学影像在不同医疗机构、设备厂商、扫描参数下存在显著的域差异。例如在某一品牌 CT 设备上训练的模型，往往在另一品牌设备的图像上性能大幅下降[91]，对抗性扰动与图像伪影也严重威胁临床安全。未来，通过联邦学习[92]在保护隐私前提下跨机构协同学习，并赋予模型持续学习的能力，使其能在不遗忘旧知识的前提下适应新数据与新设备的成像特征，是解决泛化难题的关键。

3.2 医学影像大模型在重大疾病中的临床应用与未来展望 随着大模型能力的跃升，其在各类重大疾病的筛查、诊断、治疗规划及预后评估中的应用场景正得到空前拓展。

3.2.1 肿瘤 传统的临床路径依赖专病专查，而大模型凭借其强大的高维特征表征能力，正在开启“机会性筛查”的新纪元。例如，在肿瘤的精准风险分层中，未来的研究范式可以利用自监督学习在海量常规影像数据上进行预训练。通过设定前置任务，大模型能够在无人工标注下，学习到与患者全身代谢、内分泌及组织衰老相关的深层三维空间表型。随后将这些具备强大泛化能力的基础模型微调应用于特定的下游任务，结合影像报告和临床指南标准，模型不仅能辅助发现肉眼易漏诊的微小病灶，还能实现高精度的良恶性二分类预测及长期患癌风险的动态评估。这种利用非特异性常规影像挖掘深层肿瘤生物学特征的思路，打破了传统“头痛医头”的局限，极大地提高了现有常规医疗数据的利用率。此外，在肺癌和消化道肿瘤中，大模型结合多期增强 CT 和随访数据，有望在预测肿瘤微环境异质性、基因突变状态及免疫治疗响应率方面发挥颠覆性作用。

3.2.2 神经系统疾病 在多模态融合与退行性疾病的早期预警的研究中，此类疾病（如阿尔茨海默病、帕金森病及脑卒中）通常伴随复杂的脑组织形态学与功能学改变。传统的单一 MRI 序列分析难以全面捕捉疾病早期的微观演变。医学影像大模型通过整合 T1/T2 加权成像、DTI（弥散张量成像）及 fMRI 等多模态序列，能够在三维甚至四维（时空）层面上建立全脑网络的连接拓扑图。未来的临床展望在于，视觉-语言大模型可以将患者的脑部 MRI 与量表测试得分、脑脊液生物标志物甚至基因组数据进行深度融合。通过上下文理解，大模型不仅能够自动识别海马体萎缩或白质高信号的微小变化，还能将其与患者的认知衰退轨迹对齐，从而在轻度认知障碍阶段实现精准的早期预警。对于急性脑卒中，通用模型有望在极短时间内自动量化梗死核心与缺血半暗带，并结合临床文本即时生成溶栓或取栓的收益-风险评估报告。

3.2.3 心血管与呼吸系统疾病 此类疾病需要精准量化与全周期管理。在心血管领域，超声心动图、冠脉 CTA 及心脏磁共振的解读高度依赖医师经验，且极具主观性。大模型（如应用于心肌瘢痕量化的 Scar-Net 等架构原理）能够在高背景噪声下实现对心肌、心室容积及血管斑块负荷的全自动、高精度三维重建与量化。未来，大模型有望通过单次扫描的非门控胸部 CT，同步完成冠状动脉钙化积分的自动化提取，实现冠心病风险的无感筛查。在呼吸系统疾病方面，胸部 X 线和 HRCT 是慢阻肺、间质性肺病及各类感染性肺炎的核心诊断工具。基于 MIMIC-CXR 等数据集预训练的视觉-语言模型，已经展现出自动鉴别复杂肺部感染并撰写高质量初稿的能力。未来的大模型将进一步结合患者的既往病史、吸烟史及肺功能测试结果，不仅对当前的实变或磨玻璃影进行定性诊断，还能预测肺纤维化的进展速率，辅助制定个体化的长期干预方案。

3.3 迈向医疗通用智能与全流程决策辅助 除了在具体病种上的深化，模型学习能力的进化也是关键，面对罕见病和新发疾病标注极度稀缺的现状，未来的研究将高度依赖小样本学习和零样本学习。通过上下文学习和提示工程，使模型仅凭少量示例或文本描述即可识别罕见病灶。在具备快速学习能力后，未来的医学影像大模型将不再局限于阅片室，而是向整合多源数据的“全能型医疗助手”演进。智能体模型的兴起预示着人工智能将从被动的辅助诊断工具转变为主动的协作伙伴[93]。这种多模态智能体不仅能自动抓取影像异常并调阅历史电子病历进行比对，还能自主规划复查流程、提醒临床医师潜在的漏诊风险，甚至有望在手术导航系统中与手术机器人协同工作，打通从“影像感知诊断”到“物理干预治疗”的全闭环。

3.4 破解黑箱信任危机与突破部署瓶颈 在高风险

的临床医疗中，深度学习模型通常被视为“黑箱”。缺乏可解释性不仅阻碍了医师的信任，也使得美国FDA、中国 NMPA 等机构的监管审批变得异常困难。虽然目前注意力图和显著性图技术有所进展，但往往仅提供事后解释，难以揭示深层的因果推理机制[94]。未来，结合因果推理和符号人工智能的混合模型，将促使模型不仅“知其然”（给出诊断结论），更“知其所以然”（提供符合医学逻辑的推理链条）。同时，检索增强生成技术将允许大模型在推理时动态检索最新的医学文献和临床指南，通过整合结构化医学知识图谱，为生成的诊断和治疗建议提供严谨的循证医学依据[95]。

最后，大模型动辄数十亿的参数规模对基层医院的计算资源构成了显著障碍。在保持性能的前提下，广泛应用参数高效微调（PEFT，如 LoRA 机制）和模型压缩（量化、剪枝）技术以满足临床实时推理的工作流需求，并积极解决模型可能继承的针对少数族裔或女性的诊断偏见，确保系统在真实世界临床环境中的公平性、透明性与可问责性，是未来走向大规模临床普及必须正视和解决的现实课题。

## 4 结论

医学影像大模型标志着医疗 AI 向多模态基础模型的重要范式转变，为医疗通用人工智能铺平了道路。本文系统梳理了医学影像大模型的技术架构、学习范式与临床应用，并剖析了数据质量、模型泛化及监管伦理等挑战。展望未来，医学影像大模型的发展将朝着更通用、更智能、更可信的方向迈进。通过持续学习、小样本适应、因果推理和检索增强等前沿技术的融合，下一代医学人工智能有望成为医师真正的“智能助手”，在诊断、治疗规划、预后预测等全流程中发挥关键作用。然而，技术进步必须与伦理规范、临床验证和监管框架同步推进，唯有如此，医学影像大模型才能真正惠及全球患者，推动医疗健康事业的高质量发展。

**【参考文献】**


[1] ALZUBAIDI L, ZHANG J L, HUMAIDI A J, *et al*. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions[J]. J Big Data, 2021, 8(1): 53.

[2] ARDILA D, KIRALY A P, BHARADWAJ S, *et al*. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography[J]. Nat Med, 2019, 25(6): 954-961.

[3] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[4] ACHIAM J, ADLER S, AGARWAL S, *et al*. GPT-4 technical report[EB/OL]. 2023: arXiv: 2303.08774. https://arxiv.org/abs/2303.08774

[5] BOMMASANI R, HUDSON D A, ADELI E, *et al*. On the opportunities and risks of foundation models[EB/OL]. 2021: arXiv: 2108.07258. https://arxiv.org/abs/2108.07258

[6] PAŁCZYŃSKI K, ŚMIGIEL S, LEDZIŃSKI D, *et al*. Study of the few-shot learning for ECG classification based on the PTB-XL dataset[J]. Sensors, 2022, 22(3): 904.

[7] LIU C, WAN Z W, OUYANG C, *et al*. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement [EB/OL]. 2024: arXiv: 2403.06659. https://arxiv.org/abs/2403.06659

[8] KIRILLOV A, MINTUN E, RAVI N, *et al*. Segment anything [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 3992-4003.

[9] MA J, HE Y T, LI F F, *et al*. Segment anything in medical images[J]. Nat Commun, 2024, 15(1): 654.

[10] MOOR M, BANERJEE O, ABAD Z S H, *et al*. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.

[11] WU C Y, ZHANG X M, ZHANG Y, *et al*. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data[J]. Nat Commun, 2025, 16(1): 7866.

[12] LI X, ZHAO L, ZHANG L, *et al*. Artificial general intelligence for medical imaging analysis[J]. IEEE Rev Biomed Eng, 2025, 18: 113-129.

[13] COX J, LIU P, STOLTE S E, *et al*. BrainSegFounder: towards 3D foundation models for neuroimage segmentation[J]. Med Image Anal, 2024, 97: 103301.

[14] TAVAKOLI N, ALI RAHSEPAR A, BENEFIELD B C, *et al*. ScarNet: a novel foundation model for automated myocardial scar quantification from late gadolinium-enhancement images [J]. J Cardiovasc Magn Reson, 2025, 27(2): 101945.

[15] OQUAB M, DARCET T, MOUTAKANNI T, *et al*. DINOv2: learning robust visual features without supervision [EB/OL]. 2023: arXiv: 2304.07193. https://arxiv.org/abs/2304.07193

[16] PÉREZ-GARCÍA F, SHARMA H, BOND-TAYLOR S, *et al*. Exploring scalable medical image encoders beyond text supervision[J]. Nat Mach Intell, 2025, 7(1): 119-130.

[17] MH NGUYEN D, NGUYEN H, DIEP N, *et al*. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching[J]. Advances in Neural Information Processing Systems, 2023, 36: 27922-27950.

[18] MA J, KIM S, LI F F, *et al*. Segment anything in medical images and videos: benchmark and deployment [EB/OL]. 2024: arXiv: 2408.03322. https://arxiv.org/abs/2408.03322.

[19] YANG Y Q, XU L, TIAN L X. SAM2-3dMed: empowering SAM2 for 3D medical image segmentation[EB/OL]. 2025: arXiv: 2510.08967. https://arxiv.org/abs/2510.08967

[20] GAO Y H, LIU D, LI Z W, *et al*. Show and segment: universal medical image segmentation via in-context learning [C]//

2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2025: 20830-20840.

[21] RADFORD A, KIM J W, HALLACY C, *et al*. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021.

[22] WANG Z F, WU Z B, AGARWAL D, *et al*. MedCLIP: contrastive learning from unpaired medical images and text[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab EmiratesAssociation for Computational Linguistics, 2022: 3876-3887.

[23] ZHANG S, XU Y B, USUYAMA N, *et al*. A multimodal biomedical foundation model trained from fifteen million image-text pairs[J]. Nejm Ai, 2025, 2(1): AIoa2400640.

[24] YOU K, GU J, HAM J, *et al*. CXR-CLIP: toward large scale chest X-ray language-image pre-training[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Cham: Springer, 2023: 101-111.

[25] GHOSH S, POYNTON C B, VISWESWARAN S, *et al*. Mammo-CLIP: a Vision Language Foundation Model toEnhance Data Efficiency andRobustness inMammography[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2024. Cham: Springer, 2024: 632-642.

[26] HAMAMCI I E, ER S, ALMAS F, *et al*. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities [J]. CoRR, 2024.

[27] CHRISTENSEN M, VUKADINOVIC M, YUAN N, *et al*. Vision-language foundation model for echocardiogram interpretation[J]. Nat Med, 2024, 30(5): 1481-1488.

[28] SAAB K, TU T, WENG W H, *et al*. Capabilities of gemini models in medicine[EB/OL]. 2024: arXiv: 2404.18416. https://arxiv.org/abs/2404.18416

[29] TU T, AZIZI S, DRIESS D, *et al*. Towards generalist biomedical AI[J]. Nejm Ai, 2024, 1(3): AIoa2300138.

[30] BAI F, DU Y X, HUANG T J, *et al*. M3D: advancing 3D medical image analysis with multi-modal large language models [EB/OL]. 2024: arXiv: 2404.00578. https://arxiv.org/abs/2404.00578

[31] BLANKEMEIER L, COHEN J P, KUMAR A, *et al*. Merlin: a vision language foundation model for 3D computed tomography [J]. Res Sq, 2024: rs.3.rs-rs.4546309.

[32] JAEGLE A, BORGEAUD S, ALAYRAC J B, *et al*. Perceiver IO: a general architecture for structured inputs & outputs[EB/OL]. 2021: arXiv: 2107.14795. https://arxiv.org/abs/2107.14795

[33] AKBARI H, YUAN L Z, QIAN R, *et al*. VATT: transformers for multimodal self-supervised learning from raw video, audio and text[EB/OL]. 2021: arXiv: 2104.11178. https://arxiv.org/abs/2104.11178

[34] WAXLER S, BLAZEK P, WHITE D, *et al*. Generative medical event models improve with scale[EB/OL]. 2025: arXiv: 2508.12104. https://arxiv.org/abs/2508.12104

[35] YUAN R T, ZHOU L, XU J L, *et al*. TGSAM-2: text-guided medical image segmentation using segment anything model 2 [C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2025. Cham: Springer, 2026: 565-574.

[36] ANWAR S M, MAJID M, QAYYUM A, *et al*. Medical image analysis using convolutional neural networks: a review[J]. J Med Syst, 2018, 42(11): 226.

[37] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. 2020: arXiv: 2010.11929. https://arxiv.org/abs/2010.11929

[38] LI J, CHEN J Y, TANG Y C, *et al*. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives[J]. Med Image Anal, 2023, 85: 102762.

[39] SHAMSHAD F, KHAN S, ZAMIR S W, *et al*. Transformers in medical imaging: a survey[J]. Med Image Anal, 2023, 88: 102802.

[40] LI C H, ZHANG C N. Toward a deeper understanding: RetNet viewed through Convolution[J]. Pattern Recognit, 2024, 155: 110625.

[41] KHAN A, RAUF Z, SOHAIL A, *et al*. A survey of the vision transformers and their CNN-transformer based variants[J]. Artif Intell Rev, 2023, 56(3): 2917-2970.

[42] HUANG Z Y, WANG H Y, DENG Z Y, *et al*. STU-Net: scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training[EB/OL]. 2023: arXiv: 2304.06716. https://arxiv.org/abs/2304.06716

[43] CHEN Z Y, ZHU Y S, ZHAO C Y, *et al*. DPT: deformable patch-based transformer for visual recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event China: ACM, 2021: 2899-2907.

[44] HAN K, WANG Y H, CHEN H T, *et al*. A survey on vision transformer[J]. IEEE Trans Pattern Anal Mach Intell, 2023, 45(1): 87-110.

[45] BI J R, ZHU Z L, MENG Q L. Transformer in computer vision [C]//2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). Fuzhou: IEEE, 2021: 178-188.

[46] OKOLO G I, KATSIGIANNIS S, RAMZAN N. IEViT: an enhanced vision transformer architecture for chest X-ray image classification[J]. Comput Methods Programs Biomed, 2022, 226: 107141.

[47] XIAO H G, LI L, LIU Q Y, *et al*. Context-aware and local-aware fusion with transformer for medical image segmentation [J]. Phys Med Biol, 2024, 69(2). DOI: 10.1088/1361-6560/ad14c6.

[48] ZHANG C Y, SUN S B, HU W M, *et al*. FDR-TransUNet: a novel encoder-decoder architecture with vision transformer for improved medical image segmentation[J]. Comput Biol Med, 2024, 169: 107858.

[49] WANG C, WANG L, WANG N Q, *et al*. CFATransUnet: channel-wise cross fusion attention and transformer for 2D medical image segmentation [J]. Comput Biol Med, 2024, 168: 107803.

[50] LI C L, TANG T, WANG G R, *et al*. BossNAS: exploring hybrid CNN-transformers with block-wisely self-supervised neural architecture search[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021.

[51] CHEN J N, LU Y Y, YU Q H, *et al*. TransUNet: transformers make strong encoders for medical image segmentation[EB/OL]. 2021: arXiv: 2102.04306. https://arxiv.org/abs/2102.04306

[52] LUO Y M, WANG Y, ZU C, *et al*. 3D transformer-GAN for high-quality PET reconstruction[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Cham: Springer, 2021: 276-285.

[53] LIU D, GAO Y H, ZHANGLI Q, *et al*. TransFusion: multi-view divergent fusion forMedical image segmentation withTransformers[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2022. Cham: Springer, 2022: 485-495.

[54] JIA C, YANG Y F, XIA Y, *et al*. Scaling up visual and vision-language representation learning with noisy text supervision [C]//International Conference on Machine Learning. 2021

[55] BITTO A, FÜRST A, HOCHREITER S, *et al*. CLOOB: modern Hopfield networks with InfoLOOB outperform CLIP [C]//Advances in Neural Information Processing Systems 35. New Orleans : Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 20450-20468.

[56] LI Y G, LIANG F, ZHAO L C, *et al*. Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm[EB/OL]. 2021: arXiv: 2110.05208. https://arxiv.org/abs/2110.05208

[57] ESLAMI S, MEINEL C, DE MELO G. PubMedCLIP: how much does CLIP benefit visual question answering in the medical domain? [C]//Findings of the Association for Computational Linguistics: EACL 2023. Dubrovnik, CroatiaAssociation for Computational Linguistics, 2023: 1181-1193.

[58] ZHANG Y H, JIANG H, MIURA Y, *et al*. Contrastive learning of medical visual representations from paired images and text [EB/OL]. 2020: arXiv: 2010.00747. https://arxiv.org/abs/2010.00747

[59] WANG F, ZHOU Y, WANG S, *et al*. Multi-granularity cross-modal alignment for generalized medical visual representation learning[C]//Advances in Neural Information Processing Systems 35. New Orleans: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 33536-33549.

[60] LI L H, YATSKAR M, YIN D, *et al*. VisualBERT: a simple and performant baseline for vision and language [EB/OL]. 2019: arXiv: 1908.03557. https://arxiv.org/abs/1908.03557

[61] CHEN Z H, DU Y H, HU J P, *et al*. Multi-modal masked autoencoders for medical vision-and-language pre-training[M]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2022. Cham: Springer Nature, 2022: 679-689.

[62] WANG Z R, YU J H, YU A W, *et al*. SimVLM: simple visual language model pretraining with weak supervision [EB/OL]. 2021: arXiv: 2108.10904. https://arxiv.org/abs/2108.10904

[63] CHEN J, GUO H, YI K, *et al*. VisualGPT: data-efficient adaptation of pretrained language models for image captioning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 18009-18019.

[64] ALAYRAC J B, BARR I, BARREIRA R, *et al*. Flamingo: a visual language model for few-shot learning[C]//Advances in Neural Information Processing Systems 35. New Orleans: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022: 23716-23736.

[65] GAO J F, LI C Y, LIU H T, *et al*. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day [C]//Advances in Neural Information Processing Systems 36. New Orleans: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 28541-28564.

[66] DAI W L, FUNG P N, HOI S, *et al*. InstructBLIP: towards general-purpose vision-language models with instruction tuning [C]//Advances in Neural Information Processing Systems 36. December 10-16, 2023. New Orleans: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 49250-49267.

[67] LITJENS G, KOOI T, BEJNORDI B E, *et al*. A survey on deep learning in medical image analysis[J]. Med Image Anal, 2017, 42: 60-88.

[68] CHEN T, KORNBLITH S, NOROUZI M, *et al*. A simple framework for contrastive learning of visual representations[EB/OL]. 2020: arXiv: 2002.05709. https://arxiv.org/abs/2002.05709

[69] AZIZI S, CULP L, FREYBERG J, *et al*. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging[J]. Nat Biomed Eng, 2023, 7(6): 756-779.

[70] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding [C]//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). Minneapolis, MinnesotaAssociation for Computational Linguistics, 2019: 4171-4186.

[71] KHAN A, RAUF Z, REHMAN KHAN A, *et al*. A recent survey of vision transformers for medical image segmentation [J]. IEEE Access, 2025, 13: 191824-191849.

[72] HATAMIZADEH A, NATH V, TANG Y C, *et al*. Swin UNETR: swin Transformers forSemantic Segmentation ofBrain Tumors inMRI Images[C]//Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer, 2022: 272-284.

[73] RYU J S, KANG H, CHU Y, *et al*. Vision-language foundation models for medical imaging: a review of current practices and innovations[J]. Biomed Eng Lett, 2025, 15(5): 809-830.

[74] LIN H N, XU C, QIN J. Taming vision-language models for medical image analysis: a comprehensive review [EB/OL]. 2025: arXiv: 2506.18378. https://arxiv.org/abs/2506.18378

[75] MEI X Y, LIU Z L, ROBSON P M, *et al*. RadImageNet: an open radiologic deep learning research dataset for effective trans-

fer learning[J]. Radiol Artif Intell, 2022, 4(5): e210315.

[76] IRVIN J, RAJPURKAR P, KO M, *et al*. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison[J]. Proc AAAI Conf Artif Intell, 2019, 33(1): 590-597.

[77] WANG X S, PENG Y F, LU L, *et al*. ChestX-ray: hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common Thorax diseases[M]// Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. Cham: Springer International Publishing, 2019: : 369-392.

[78] JEONG J J, VEY B L, BHIMIREDDY A, *et al*. The EMory BrEast imaging dataset (EMBED): a racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images[J]. Radiol Artif Intell, 2023, 5(1): e220047.

[79] HALLING-BROWN M D, WARREN L M, WARD D, *et al*. OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data[J]. Radiol Artif Intell, 2020, 3(1): e200103.

[80] FLANDERS A E, PREVEDELLO L M, SHIH G, *et al*. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge[J]. Radiol Artif Intell, 2020, 2(3): e190211.

[81] CALLEJAS M F, LIN H M, HOWARD T, *et al*. Augmentation of the RSNA pulmonary embolism CT dataset with bounding box annotations and anatomic localization of pulmonary emboli[J]. Radiol Artif Intell, 2023, 5(3): e230001.

[82] LIN H M, COLAK E, RICHARDS T, *et al*. The RSNA cervical spine fracture CT dataset[J]. Radiol Artif Intell, 2023, 5(5): e230034.

[83] YE J, CHENG J L, CHEN J P, *et al*. SA-Med2D-20M dataset: segment anything in 2D medical imaging with 20 million masks [EB/OL]. arXiv E Prints, 2023: arXiv: 2311.11969.

[84] WASSERTHAL J, BREIT H C, MEYER M T, *et al*. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images[J]. Radiol Artif Intell, 2023, 5(5): e230024.

[85] MA J, ZHANG Y, GU S, *et al*. AbdomenCT-1K: is abdominal organ segmentation a solved problem? [J]. IEEE Trans Pattern Anal Mach Intell, 2022, 44(10): 6695-6714.

[86] JOHNSON A E W, POLLARD T J, BERKOWITZ S J, *et al*. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports[J]. Sci Data, 2019, 6(1): 317.

[87] FENG S J, AZZOLLINI D, KIM J S, *et al*. Curation of the CANDID-PTX dataset with free-text reports[J]. Radiol Artif Intell, 2021, 3(6): e210136.

[88] LAU J J, GAYEN S, BEN ABACHA A, *et al*. A dataset of clinically generated visual questions and answers about radiology images[J]. Sci Data, 2018, 5: 180251.

[89] LIU B, ZHAN L M, XU L, *et al*. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). Nice: IEEE, 2021: 1650-1654.

[90] LONGPRE S, MAHARI R, CHEN A, *et al*. A large-scale audit of dataset licensing and attribution in AI[J]. Nat Mach Intell, 2024, 6(8): 975-987.

[91] MCKINNEY S M, SIENIEK M, GODBOLE V, *et al*. International evaluation of an AI system for breast cancer screening[J]. Nature, 2020, 577(7788): 89-94.

[92] RIEKE N, HANCOX J, LI W Q, *et al*. The future of digital health with federated learning[J]. NPJ Digit Med, 2020, 3: 119.

[93] MORITZ M, TOPOL E, RAJPURKAR P. Coordinated AI agents for advancing healthcare[J]. Nat Biomed Eng, 2025, 9(4): 432-438.

[94] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [J]. Nat Mach Intell, 2019, 1(5): 206-215.

[95] LEWIS P, PEREZ E, PIKTUS A, *et al*. Retrieval-augmented generation for knowledge-intensive NLP tasks[EB/OL]. 2020: arXiv: 2005.11401. https://arxiv.org/abs/2005.11401

**（收稿日期：2026-02-15；修回日期：2026-03-10；编辑：黎仕娟）**