·智能眼科·

通用和眼科专用大模型在眼科图像中的应用效果及局限

赖世欣,范心妍,罗明杰

引用:赖世欣,范心妍,罗明杰. 通用和眼科专用大模型在眼科图像中的应用效果及局限. 国际眼科杂志, 2026,26(9):1651-1657.

基金项目:广东省自然科学基金面上项目(No.2024A1515012292);中山大学中山眼科中心优秀科技创新人才支持与培育计划项目(No.CXQN-010)

作者单位:(510623)中国广东省广州市,中山大学中山眼科中心 眼病防治全国重点实验室 广东省眼科视觉科学重点实验室
作者简介:赖世欣,本科,研究方向:医学人工智能的眼科临床评价。
通讯作者:罗明杰,博士,副研究员,研究方向:眼科医学大数据及生成式人工智能诊疗研发与应用. luomingjie@gzzoc.com
收稿日期:2026-02-06 修回日期:2026-07-23

摘要

随着大模型的发展,通用和眼科专用大模型在眼科图像分析中展现出巨大潜力,但其在特定临床任务中的应用效能及适用边界尚待明确。文章梳理了眼科大模型的技术路线演进,包括掩码视觉建模、图文对比学习以及知识融合与多模态推理等阶段,并对比分析了通用模型与专用模型在疾病筛查与初诊、疾病分级与病灶分割、医疗报告生成与多模态推理等常见临床场景中的性能差异。结果显示,通用模型在多病种初筛中表现较好,而眼科专用模型在精细化分级、微小病灶识别及罕见病诊断等任务中占据优势,且经指令微调后生成的医疗报告更具临床一致性,这些结果为具体临床任务中的模型类型及技术路线选择提供了参考。此外,文章进一步指出了当前模型面临的可靠性、泛化性、评估体系及可重复性挑战及其所造成的临床应用局限性。

关键词:视觉语言模型;人工智能;眼科;应用效果;评估
DOI:10.3980/j.issn.1672-5123.2026.9.24

Application performance and limitations of general and ophthalmology-specific large models in ophthalmic images

Lai Shixin, Fan Xinyan, Luo Mingjie

Foundation items: General Program of the Guangdong Provincial Natural Science Foundation (No.2024A1515012292); Support and Cultivation Program for Outstanding Scientific and Technological Innovation Talents, Zhongshan Ophthalmic Center, Sun Yat-sen University (No.CXQN-010)

State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University; Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510623, Guangdong Province, China

Correspondence to: Luo Mingjie. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University; Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510623, Guangdong Province, China. luomingjie@gzzoc.com
Received:2026-02-06 Accepted:2026-07-23

Abstract

• With the rapid advancement of large language models, both general-purpose and ophthalmology-specific variants have shown substantial potential for ophthalmic image analysis. However, their task-specific effectiveness and applicable boundaries in clinical practice remain to be fully clarified. This review outlines the technical evolution of large models in ophthalmology, including masked visual modeling, vision-language contrastive learning, and subsequent stages of knowledge integration and multimodal reasoning. The performance of general-purpose versus ophthalmology-specific models across common clinical scenarios, including disease screening and initial triage, disease grading and lesion segmentation, and medical report generation with multimodal reasoning were further compared. Results show that general-purpose models are well-suited for multi-disease screening and primary-care triage, whereas ophthalmology-specific models excel in fine-grained grading, subtle lesions detection, and rare disease diagnosis. Moreover, instruction-tuned domain-specific models produce medical reports with greater clinical consistency. These findings provide practical guidance for selecting appropriate model types and technical pathways tailored to specific ophthalmic tasks. Furthermore, the review identifies the current challenges facing these models, including reliability, generalizability, evaluation frameworks, and reproducibility, as well as the resulting limitations in clinical application.

• KEYWORDS: vision-language models; artificial intelligence; ophthalmology; application performance; evaluation

Citation: Lai SX, Fan XY, Luo MJ. Application performance and limitations of general and ophthalmology-specific large models in ophthalmic images. Guoji Yanke Zazhi(Int Eye Sci), 2026,26(9):1651-1657.

0 引言

人工智能(artificial intelligence, AI)技术的发展正在改变着眼科诊疗模式。早期研究主要集中于卷积神经网络(convolutional neural networks, CNN)在单一模态图像上的特定任务识别,如糖尿病视网膜病变筛查、青光眼检测

等[1]。然而,传统深度学习模型存在对标注数据依赖性强、泛化能力弱及缺乏语义理解等局限。目前应用于眼科图像的模型按照技术路线来分主要包括具有强大特征提取能力的基础模型(foundation model, FM)[2]和具备跨模态理解能力的视觉语言模型(vision-language model, VLM)[3]。按照训练数据进行分类,一类是基于通用图文对训练的 GPT-4V[4]、Gemini[5]等通用视觉语言模型,另一类是在预训练或微调阶段使用眼科数据集学习眼科影像特征的眼科专用模型如 RETFound[2]、EyeCLIP[6]。两种模型在不同临床应用场景中各有优劣。现有综述已从不同角度总结了大语言模型和基础模型在眼科中的应用进展,包括大语言模型在眼科应用场景、应用效果、发文趋势和研究图谱[7-8],基础模型、视觉和视觉语言模型的技术架构及其在眼病诊断中的准确性[9-10];或从具体眼科疾病出发,总结人工智能技术的应用和挑战[11-13],但仍缺乏综述系统梳理这两类模型的技术演进逻辑,并从临床实际需求出发,归纳其在不同诊疗环节中的表现差异。因此,本文回顾了眼科大模型的技术路线演进,从掩码视觉建模到多模态推理增强和交互梳理其发展脉络,并基于临床应用场景,对比通用模型与专用模型在疾病诊断分级、病灶分割及医疗报告生成等常见任务中的性能,剖析了现阶段模型在可靠性、泛化性及评估体系方面面临的挑战,为具体临床任务中的模型等技术路线选择提供参考。

通用视觉语言模型和眼科专用模型的代表性技术路线与临床应用归纳见表 1。本表为基于不同文献的叙述性归纳,旨在概括通用视觉语言模型与眼科专用模型的技术路线、训练数据类型及潜在的优势临床应用场景。各模型来自不同研究,其训练数据规模、数据类型、任务设置、提示词、测试集和评价指标等不完全一致,因此不能据此直接比较模型性能优劣。

1 眼科大模型的技术路线演进

眼科大模型的技术围绕数据效率、语义理解、逻辑推理和交互的问题,不断迭代更新。

1.1 掩码视觉建模与自监督学习 为解决眼科医疗场景中高质量标注稀缺的痛点,第一阶段的技术演进聚焦于掩码视觉建模,模型自监督学习未标注影像的特征。

RETFound 是首个专门针对眼科图像进行预训练的视觉基础模型,采用掩码自编码器(masked autoencoder, MAE)进行自监督预训练,模型通过“掩码-重建”学习视网膜的特征,显著降低了模型对标注数据的依赖[2]。在青光眼的诊断中,RETFound 仅需 5 轮训练(2 000 张图像)就能达到与使用了 10 000 张图像训练的 ResNet-50 相同的效果,受试者工作特征曲线下面积(area under the receiver operating characteristic curve, AUC)为 0.766[17]。在此基础上,VisionFM 采用来自多中心的多模态眼科图像进行训练,在多种任务、不同人群和场景中展现出较好的泛化性[18]。不同于 RETFound 关注像素级特征还原,MIRAGE 利用自动化算法生成的视网膜分层标记,使得模型学习视网膜解剖特征,在 OCT 和 SLO 图像的分类分割任务中表现出更好的性能[19]。

1.2 图文对比学习与语义理解 尽管视觉模型在眼科图像特征提取中表现良好,但对图像的语义理解仍有限。图文对比学习通过在大规模图文对上拉近图像与文本特征的距离,使得模型获得“零样本”推理能力,进一步提升了模型对数据的利用效率。

RET-CLIP 在 CLIP 的基础上引入基于变换器的双向编码器(bidirectional encoder representation from transformers, BERT)的文本编码器,增强了模型学习医疗报告上下文的能力[14]。针对公开眼科数据集普遍存在文本描述匮乏或含有噪声的问题,KeepFIT 和 KeepFIT V2 采取了“图像相似性引导的文本修正”的策略,通过计算图像相似度,将 MM-Retinal 中的高质量专家描述“迁移”到公开数据集中视觉特征相似的图像上,从而显著丰富了公开数据的语义信息并修正了潜在噪声[20]。ViLReF 在传统对比学习的基础上,引入了专家知识作为监督信号,并提出加权相似度耦合损失(weighted similarity coupling Loss),以降低语义相同却被误判为负例的假阴性率[21]。

1.3 知识融合与混合训练 掩码建模自监督学习与图文对比学习的结合使得模型兼具图像特征学习能力与语义理解能力。UrFound 提出了知识引导掩码建模(knowledge-

表 1 通用视觉语言模型和眼科专用模型的代表性技术路线与临床应用归纳

比较维度	通用视觉语言模型	眼科专用模型
技术路线及其代表模型(训练数据)	多模态预训练与指令/偏好对齐:GPT-4V[4]、Gemini 2.5[5](通用图文对;数据规模未披露)	基于眼科图像的自监督视觉建模:RETFound[2](90 万张眼底图像,73 万张 OCT 图像) 眼科图文对比学习:RET-CLIP[14](19 万患者的眼底彩照及临床诊断报告) 掩码视觉建模与图文对比学习结合: EyeCLIP[6](277 万张 11 种模态眼科图像及 1 万份临床报告) 视觉编码器与大语言模型结合:EyeFM[15](1 450 万张 5 种模态眼科图像及 40 万份临床报告) 微调通用视觉语言模型:FundusExpert[16](20 万张眼底图像,30 万条图文指令微调样本)
优势能力	多模态推理、对话交互、语义理解	特征提取、病灶分割、罕见病识别
临床局限	易产生幻觉、微小病灶识别能力较弱	跨模态推理能力受限、泛化性不足
适用场景	常见病初筛、患者对话交互	罕见病和复杂病例识别、疾病分级、病灶定位和分割、专业医疗报告生成

guided masked modeling strategies)策略,将文本知识注入到视觉特征的学习过程中,增强了模型对病灶的理解,在多模态联合训练中证明了其互补性[22]。类似地,EyeCLIP融合了自监督训练、多模态图像对比学习和图文预训练,在涵盖11种模态的超大规模数据上进行了统一训练,从而在零样本学习和跨模态识别任务中均取得了超越单一范式模型的表现[6]。针对纯图文对比学习模型难以捕捉微小病灶轮廓的局限,RetiZero融合了MAE与CLIP,再引入"图-图检索(Image-to-Image Retrieval)"策略,在无需重新训练的情况下,通过检索知识库中视觉特征相似的确诊病例,即可较为精准识别包括罕见病在内的400多种眼底疾病,Top-5准确率为0.756[23]。

1.4 交互与推理增强 随着多模态大模型的出现,眼科模型通过连接视觉编码器与LLM或微调通用VLM,模拟医生的临床思维过程,进行疾病分析和医疗报告生成。在架构构建方面,OphGLM[24]、EyecareGPT[25]和EyeFM[15]通过融合模块将视觉特征映射到LLM的特征空间,实现了基于图像的对话交互和报告生成;RetinaVLM则通过两阶段训练(base-specialist)连接ResNet50与Llama3,强化了对生物标志物与疾病分期的关联分析能力[26]。此外,针对眼科超声高度依赖专家解读的痛点,视觉语言分割(vision-language segmentation, VLS)在实现病灶精准分割的同时生成诊断报告,显著提升了模型在超声影像分析中的可解释性[27]。另一类模型则采用通用VLM微调,如FundusExpert引入"认知链(cognitive chain)"设计,模拟从局部病灶观察到全局诊断的推理过程[16]。

2 眼科大模型在关键临床场景下的应用效果评估

我们从疾病筛查与初诊、精细化分级与病灶分割、医疗报告生成与多模态推理三个常见临床场景出发,梳理通用视觉语言模型(表2)与眼科专用模型(表3)的应用效果。表中结果来自不同研究,其数据集来源、样本量、任务定义、模型版本、提示词设置、训练/微调方式及评价指标均不完全一致,因此不宜据此直接比较不同模型的性能优劣。

2.1 疾病筛查与初诊 在常见的二分类筛查任务中,通用视觉语言模型凭借广泛的语义常识展现出一定潜力。研究显示,在糖尿病视网膜病变(DR)二分类任务中,ChatGPT-4o的准确率为0.64,且不同通用多模态大模型的特异度均较高(0.87-0.99),提示其在DR识别中具有一定潜力,但整体性能仍不足以支持安全的临床应用[28]。然而,通用模型在多眼病分类任务中的表现呈现出不稳定性与异质性。一项研究显示,GPT-4V诊断6种常见眼病的准确率高达97.1%,显著优于Gemini(41.2%)[29];而在另一项研究中,Gemini的准确率(61%)高于GPT-4(51%)和Copilot(29%)[30]。这些研究结论的异质性表明由于缺乏统一的测试基准(benchmark)和标准化的评估框架,模型性能极易受数据集分布、提示词工程(prompt engineering)及任务设置的影响,导致其临床泛化能力难以被客观量化。此外,在面对罕见病或复杂眼底疾病时,通用模型体现出一定的局限性,GPT-4-turbo在图像结合

表2 通用视觉语言模型在关键眼科临床场景中的代表性研究结果

临床应用场景	通用VLM		
	模型	指标及结果	数据集
常见眼病诊断	Gemini-2.0-Flash	准确度(accuracy,ACC)=0.900[34]	MuReD,2 451张眼底彩照,20种常见眼病
疾病严重程度分级(DR分级)	ChatGPT-4o	ACC=0.395[28]	309眼的超广角眼底图像
罕见病和复杂病例诊断	GPT-4-turbo	ACC=0.493[31]	JAMA Ophthalmology's Clinical Challenges,422个病例,图像及其文本描述
医疗报告生成	GPT-4	2名视网膜亚专科医生根据"医疗正确性"和"错误严重性"对每份报告打分,均值为2.95/5.00[35]	100张OCT图像

表3 眼科专用模型在关键眼科临床场景中的代表性研究结果

临床应用场景	眼科专用大模型		
	模型	指标及结果	数据集
常见眼病诊断	RetiZero	ACC=0.442[23]	EYE-15,30 089张眼底彩照,15种常见眼病
疾病严重程度分级(DR分级)	RETFound	AUC=0.822[2]	IDRiD
	RET-CLIP	AUC=0.863[14]	
	EyeCLIP	AUC=0.757[6]	
	EyeFM	AUC=0.811[15]	
罕见病和复杂病例诊断	RetiZero	ACC=0.726[23]	包含Bietti晶体营养不良、脉络膜肿瘤等罕见病的EYE-52数据集,7 007张眼底彩照
医疗报告生成	RetinaVLM-Specialist	2名高年资眼科医生评估为"观察与结论正确"的报告占比:64.3%[26]	28张OCT图像

相应文本描述时准确率为 0.493,与眼科专家存在差距[31]。专用模型则探索了针对罕见病数据稀缺问题的特定技术路径。RetiZero 模型通过图-图检索技术在包含 52 种眼底疾病(含罕见病)的 EYE-52 数据集中实现了零样本诊断,Top-5 准确率达 0.756[23]。这表明通用模型在复杂病例中存在局限性,而专用模型则通过特定机制为罕见病和复杂病例的识别提供了可行的技术方案。

2.2 疾病分级与病灶分割 在涉及细微病灶识别的精细化任务中,通用模型表现出局限性。在 DR 严重程度分级任务中,ChatGPT-4o、Claude 3.5 Sonnet、Gemini 1.5 Pro 和 Perplexity Llama 3.1 的准确率分别为 0.395、0.314、0.392 和 0.411,且各模型敏感度均低于 0.25,提示通用多模态大模型对于疾病精细化分级的能力仍有限[28]。而经过大规模眼科影像预训练的专用基础模型表现出一定优势。研究表明,RETFound 对轻度青光眼的检测 AUC 平均值达到 0.81,中重度青光眼为 0.95[17]。VisionFM 在血管分割和区域分割任务中 Dice 相似系数分别为 81.75%和 42.07%,优于 U-Net 的表现(81.50%和 26.87%)[18]。对于定位任务,尽管 FundusExpert 能够较为准确定位视杯和视盘(IoU 分别为 0.632 和 0.738),但定位硬性渗出、棉绒斑、微血管瘤等病灶的 IoU 均不足 0.200,提示高精度的病灶分割仍需依赖具备强视觉特征提取能力的专用模型[16]。

2.3 医疗报告生成与交互推理 在医疗报告生成与临床交互推理方面,眼科专用模型通过引入特有的推理机制,显著修正了通用模型在专业性与逻辑性上的缺陷。通用模型在迭代过程中医学问答能力增强,最新的 OpenAI o1 模型在回答眼科患者咨询中的“可读性”(12.6/15)和“准确性”(14.2/15)均优于 GPT-4[32]。对 GPT-4o、Claude 3.5等 7 种通用模型的最新评估也证实,Claude 3.5 Sonnet 在纯文本眼科考试中准确率达 77.7%,接近高年资眼科医生(P=0.72);然而对于包含影像的临床题目,表现最好的通用模型 GPT-4o 准确率仅为 57.5%,显著低于眼科住院医师(71.2%)和专家(75.7%)的水平,显示模型在多模态推理上的局限性[33]。在医疗报告生成能力上,研究表明 GPT-4V(默认模式)和 Gemini(2023.10 版本)生成的“高质量”图像描述分别仅占 21.2%和 28.6%[29]。而经过指令微调的专用模型显著提升了报告的临床一致性与准确度。例如,RetinaVLM 生成的 ARMD 诊断报告准确性为64.3%,显著高于 GPT-4o 的 14.3%[26];FundusExpert 生成报告的临床一致性为 77%,比 GPT-4o 高出 29.4%[16];EyeFM 在真实世界研究中辅助医生将报告撰写时间平均缩短 63.3 秒,提升了临床效率[15]。在眼科超声领域,VLS 模型在多中心测试的报告生成任务中,外部测试集的 BLEU-4 评分达到 85.36,优于传统模型[27]。这些模型通过引入模拟医生思维的复杂推理机制,在多模态推理任务和医疗报告生成中体现出更优性能。

3 挑战和解决思路

3.1 模型准确性 现有模型在生成回答时,常产生“幻觉”,即出现与图像无关或错误的结论[36]。这些输出往往带有完整的推理链,具有一定诱导性,对临床应用造成潜在风险。根据任务类型和错误发生环节可以将眼科图文任务中出现的幻觉分为两类:(1)视觉理解幻觉,包括病灶数量、类别、位置、诊断类型等错误,例如,在视网膜前膜病例中,ChatGPT 将其误诊为分支视网膜静脉阻塞,并描述了图像中并不存在的静脉扩张和迂曲等特征[36];(2)逻辑组合幻觉,涉及模型在整合影像特征、病史和眼科知识时形成不恰当的推理链,例如模型将识别出的病理特征关联到错误的疾病或病理状态,或缺乏对疾病进展的动态推断,导致了错误的决策[37]。针对这一问题,已有方法包括:在构建模型时引入人工评估或提示词微调模块,提高输出准确率;利用检索增强生成(RAG),通过接入 PubMed 等专业文献提升内容可靠性[38, 39]。在眼科临床问答任务中,采用全面眼科数据集增强后的 ChatZOC 相比于基线模型(Baichuan-13B),其科学一致性提升了 33.5%、“无不当或错误内容”的比例提升了 42.8%,表明知识增强有助于提升模型在问答任务中的准确率[40]。然而,现有方法仍有一定局限性。指令微调高度依赖高质量的眼科图文对、专家标注和微调任务设定,在眼科落地时可能面临高质量眼科图文数据不足、专家标注成本高、不同模态和疾病任务难以统一、跨中心外部泛化不确定等问题[24, 26]。检索增强生成虽然可通过接入最新指南和专业文献提升回答准确性,但其在眼科落地仍受限于检索与生成带来的计算延迟和资源需求、检索信息质量与时效性不稳定、持续专家监督需求,以及检索来源中潜在的种族、性别等偏倚对临床建议公平性的影响[41]。因此,未来眼科大模型应结合自适应和领域特异性检索机制、高质量多模态眼科数据、专家反馈对齐、多中心外部验证和标准化幻觉评估体系,以提升模型输出的准确性、可追溯性和临床安全性。

3.2 模型泛化性 眼科图像在不同模态、成像设备及种族人群间存在显著差异,而现有模型往往依赖单一来源或有限数据进行训练,难以在跨人群、跨设备的场景中保持稳定性能。例如,RETFound 评估结果显示,当应用在亚洲人群或不同年龄组时,其分类表现与未经眼科图像训练的 ViT 模型相近,表明其泛化能力不足[42]。这反映出眼科专用基础模型更擅长学习特定眼科影像特征,但若训练数据集中于有限人群、设备或疾病谱,其专科表征也可能受训练分布限制;相反,通用基础模型虽然缺乏眼科细粒度病灶知识,但具有更广泛的视觉和语义先验,在部分跨场景任务中可能保留一定迁移能力。除了训练数据的多样性,算法层面的局限性也可能导致模型泛化性不足。医学图像中的领域偏移包括协变量偏移和概念偏移:前者主要由成像设备、采集方式、图像质量、患者人群等差异引起,后者往往来自标注者差异、诊断标准或疾病定义不一致,导致模型学到的图像-标签关系难以直接迁移到外部中心[43]。此外,模型架构本身的局限性可能导致泛化性不足,研究表明,ImageNet 预训练的 CNN 更容易关注图像的纹理而不是全局特征,提示在眼科图像中,该类模型可能依赖颜色、亮度、对比度或设备特异性图像风格,而没有学习到图像的解剖和病理结构,导致其在跨设备、跨中心测试时可能出现泛化能力下降[44]。因此,提升眼科模型泛化性应当从数据和算法两方面进行优化:(1)纳入多中心、多种族、多模态和多设备数据,覆盖真实临床中的图像质量、疾病谱和人群差异;(2)通过领域泛化、不变特征学习、多尺度特征提取等方法,减少模型对设备风格、图像颜色、亮度和质量等捷径特征的依赖。基于多中心、多种族、多模态和大规模数据训练的 EyeFM 模型对于新模态、新设备、多任务场景中泛化性良好,并在真实世界研究中显

著提升了医生的诊断准确率(92.2% *vs* 75.4%)[15],提示数据多样性与规模化预训练策略有助于增强模型在复杂临床环境下的泛化和迁移能力。未来的模型架构可探索通用基础模型与眼科专科适配模块的融合路径,在广泛医学知识、语言交互和常见病识别能力的基础上,通过眼科图像编码器、病灶定位模块、专科知识库等方法提升模型微小病灶识别、罕见病诊断和复杂分级任务的性能。

3.3 标准化评估体系

3.3.1 评估指标 现有图像分类任务和视觉问答任务主要使用 ACC、AUC/AUROC、F1 值、Top-5 准确率等指标;其中,ACC 反映模型总体判断正确的比例,AUC 反映模型在不同阈值下区分阳性与阴性病例的能力,F1 值综合考虑精确率与召回率,较适用于类别不平衡任务,Top-5 准确率可用于评估模型是否能将正确疾病纳入候选诊断范围;病灶定位及分割任务常使用交并比(intersection over union, IoU),用于衡量模型预测区域与专家标注区域的空间重叠程度,反映模型对病灶或解剖结构定位的准确性;图文生成任务则常依赖 BLEU、ROUGE、CIDEr 等衡量语言流畅度的指标,BLEU 主要基于 n-gram 词组重叠评价文本相似性,ROUGE 侧重评价关键信息召回,CIDEr 则衡量生成描述与人工参考描述之间的加权一致性[45-46]。然而,这些指标难以反映临床真实价值和生成报告的语义准确性。而目前借助 LLM 进行自动评估还存在信度和效度不足的问题,模型容易出现混淆评估维度、认知偏倚、注意力偏倚等问题,且由于缺乏专业知识和新近知识,导致模型对专业领域的评估存在局限性[47]。未来研究应由临床医生主导制定统一的评估标准,明确模型需关注的内容,以更好衡量生成结果的临床适用性。例如 EyeFM 在报告生成评估中,以来自多中心的 12 名眼科医生组成的专家组撰写的参考报告为标准,根据模型生成报告中的无关信息、事实错误、信息缺失与潜在伤害对报告进行质量分级;在随机对照试验中进一步依据临床指南对报告中的管理建议从完整性、正确性、必要性与安全性四个维度进行 Likert 评分,形成标准化总分,并评估评审者一致性及与患者依从性的关联,从而更贴近临床真实价值[15]。

3.3.2 评估基准 目前对视觉语言模型在眼科图像应用中的评估主要使用 IDRiD 等开源数据集或医学考试,这些基准往往局限于单病种或单模态,难以全面反映模型的跨模态、跨中心能力[48]。评估任务则主要集中在对眼病的识别和严重程度分级,然而,眼科生成式 AI 的关键挑战之一在于多模态能力,即融合多模态信息进行综合推理[49]。目前的通用大模型主要展示了记忆和理解层面的能力,但在涉及分析、评估的高阶临床推理(如整合病史与影像进行诊断辩证)方面仍显不足[50]。因此,未来的评估基准需要在任务与数据两方面同步升级。在任务方面,应设计能够分层考察模型认知深度的临床情境题和多模态融合推理任务。例如 MM-Retinal-Reason 增加了包含患者基本情况、病史和多种检查的病例分析题,以评估模型对于真实病例的推理分析能力[20]。这类任务应当侧重完整临床推理过程的评估,包括病史与检查信息整合是否合理、图像证据与诊断结论是否一致、跨模态信息之间是否相互支持、推理链是否完整、决策建议是否符合指南规范等。在数据方面,应当构建涵盖多模态、多病种、多任务、跨语言及跨种族人群的综合性眼科数据集,以全面评估模型的专科性能。

3.4 可重复性 目前,眼科领域大多数视觉语言模型未提供公开测试端口、测试数据和任务,且评价指标不一致,导致研究结论差异较大、可重复性不足。尽管 GPT-4o 等通用模型提供了交互界面和 API,但是由于输入指令、测试数据及人工判读标准存在差异,导致测试结果不一,难以反映模型的真实表现。Tripod-LLM 指南强调模型研究应全面透明地报告数据来源、模型版本、提示工程方法、推理设置和评估环境,从而提升结果的可解释性和可重复性[51]。这类标准化的报告框架有助于减少因测试方式不同带来的结论分歧,为眼科模型的标准化测试奠定基础。在临床落地层面,还需建立医院、科研机构和企业共同参与的产学研医协作模式。医院和眼科专家应负责提出真实临床需求、制定标注规范和判读标准;科研机构负责模型架构设计、算法优化和评估体系开发;企业则可承担数据治理、隐私保护、系统部署、版本管理和持续监测。通过真实世界研究、医生实时反馈和不良事件监测,形成从数据构建、模型训练、外部验证到临床部署的转化路径[51-53]。

4 小结

本文综述了眼科专用模型的技术路线并对比了通用视觉语言模型和眼科专用模型的临床应用表现。两类模型具有不同的适用场景:通用模型在多病种初筛与医患交互中具备优势,而眼科专用模型在精细分级、罕见病识别与报告生成等任务中展现出更高准确性。在性能差异上,通用模型在开放式对话和常见病初步判断和初筛中表现良好,但在细粒度病灶识别、多模态推理和专业报告生成中仍易出现准确性不足及幻觉;眼科专用模型则在病灶分割、罕见病识别等特定任务上表现更优,但仍有依赖大规模专科数据、泛化性不佳等局限性。当前眼科大模型临床转化的核心瓶颈在于准确性、泛化性、标准化评估体系、可重复性不足。因此,未来研究应重点围绕以下方向推进:(1)构建统一的眼科模型评估框架与标准化测试基准,覆盖多病种、多模态、多任务与跨中心场景,并明确任务依赖的临床关键指标;(2)推动临床专家牵头制定评估共识或指南,建立可复现的评价流程;(3)鼓励通用模型与专用模型融合路线探索,提高跨任务能力与可解释性。

利益冲突声明:本文不存在利益冲突。
作者贡献声明:赖世欣文献收集与整理,论文初稿撰写与修改;范心妍论文修改;罗明杰指导写作思路,论文审定与修改。所有作者阅读并同意最终的文本。

参考文献

[1] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA, 2017,318(22):2211-2223.
[2] Zhou YK, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature, 2023, 622(7981):156-163.
[3] Zhang JY, Huang JX, Jin S, et al. Vision-language models for vision tasks: a survey. IEEE Trans Pattern Anal Mach Intell, 2024, 46(8):5625-5644.
[4] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv, 2023.
[5] Comanici G, Bieber E, Schaekermann M, et al. Gemini 2.5:

pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv, 2025.
[6] Shi DL, Zhang WY, Yang JC, et al. A multimodal visual-language foundation model for computational ophthalmology. NPJ Digit Med, 2025,8:381.
[7] Shen R, Lee ESH, Hu X, et al. Large language models in ophthalmology: a bibliometric analysis. Eye Science, 2025, 2(3): 222-237.
[8] 章文成, 佘凌冰, 何晶晶. 大型语言模型在眼科中的应用.眼科学报, 2025,40(3):283-294.
[9] Sevgi M, Ruffell E, Antaki F, et al. Foundation models in ophthalmology: opportunities and challenges. Curr OpinOphthalmol, 2025,36(1):90-98.
[10] Jin K, Yu T, Ying GS, et al. A systematic review of vision and vision-language foundation models in ophthalmology. Adv Ophthalmol Pract Res, 2026,6(1):8-19.
[11] Lim JI, Rachitskaya AV, Hallak JA, et al. Artificial intelligence for retinal diseases. Asia Pac J Ophthalmol (Phila), 2024, 13(4): 100096.
[12] Issa M, Sukkarieh G, Gallardo M, et al. Applications of artificial intelligence to inherited retinal diseases: a systematic review. Surv Ophthalmol, 2025, 70(2):255-264.
[13] 郭东琳, 许发宝, 巩亚军, 等. 人工智能在眼底影像分析中的研究进展及应用现状. 眼科学报, 2022,37(3):185-193.
[14] Du JW, Guo J, Zhang WH, et al. RET-CLIP: a retinal image foundation model pre-trained with Clinical diagnostic reports. MICCAI 2024, 2024:709-719.
[15] Wu YL, Qian B, Li TY, et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. Nat Med, 2025, 31(10):3404-3413.
[16] Liu XY, Song DP. Constructing ophthalmic MLLM for positioning-diagnosis collaboration through clinical cognitive chain reasoning. IEEE, 2025: 21547-21556.
[17] Chuter B, Huynh J, Hallaj S, et al. Evaluating a foundation artificial intelligence model for glaucoma detection using color fundus photographs. Ophthalmol Sci, 2025,5(1):100623.
[18] Qiu JN, Wu J, Wei H, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. Nejm Ai, 2024,1(12).
[19] Morano J, Fazekas B, Sükei E, et al. Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. NPJ Digit Med, 2025,8:576.
[20] Wu RQ, Zhang CR, Zhang JL, et al. MM-retinal: knowledge-enhanced foundational pretraining with Fundus image-text expertise. MICCAI 2024, 2024:722-732.
[21] Yang SZ, Du JW, Guo J, et al. ViLReF: an expert knowledge enabled vision-language retinal foundation model. arXiv, 2024.
[22] Yu K, Zhou Y, Bai Y, et al. UrFound: towards universal retinal foundation models via Knowledge-guided masked modeling. MICCAI 2024, 2024:753-762.
[23] Wang M, Lin T, Lin AD, et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. Nat Commun, 2025,16:5528.
[24] Deng Z, Gao WH, Chen CC, et al. OphGLM: an ophthalmology large language-and-vision assistant. Artif Intell Med, 2024, 157:103001.
[25] Li SJ, Lin TW, Lin LS, et al. EyecareGPT: boosting comprehensive ophthalmologyunderstanding with tailored dataset, benchmark and model. USAACM, 2025: 3893-3902.
[26] Holland R, Taylor TRP, Holmes C, et al. Specialized curricula for training vision language models in retinal image analysis. NPJ Digit Med, 2025,8(1):532.
[27] Jin K, Sun QX, Kang DH, et al. Grounded report generation for enhancing ophthalmic ultrasound interpretation using Vision-Language Segmentation models. NPJ Digit Med, 2026,9:99.
[28] Most JA, Walker EH, Mehta NN, et al. Can multimodal large language models diagnose diabetic retinopathy from fundus photos? A quantitative evaluation. Ophthalmol Sci, 2026,6(1):100911.
[29] Srinivasan S, Ji HW, Chen DZ, et al. Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? BMJ Open Ophthalmol, 2025,10(1):e002076.
[30] Mihalache A, Huang RS, Popovic MM, et al. Fundus photograph interpretation of common retinal disorders by artificial intelligence chatbots. AJO Int, 2025,2(3):100154.
[31] Mikhail D, Milad D, Antaki F, et al. Multimodal performance of GPT-4 in complex ophthalmology cases. J Pers Med, 2025,15(4):160.
[32] Pushpanathan K, Zou MJ, Srinivasan S, et al. Can OpenAI's new o1 model outperform its predecessors in common eye care queries? Ophthalmol Sci, 2025,5(4):100745.
[33] Rocha H, Chong YJ, Thirunavukarasu AJ, et al. Performance of foundation models vs physicians in textual and multimodal ophthalmological questions. JAMA Ophthalmol, 2026,144(1):5-13.
[34] Liang XY, Bian MX, Chen MX, et al. A novel ophthalmic benchmark for evaluating multimodal large language models with fundus photographs and OCT images. arXiv, 2025.
[35] Chen XJ, Fu HZ, Wang JT, et al. A deep learning based automatic report generator for retinal optical coherence tomography images. NPJ Digit Med, 2025,8:618.
[36] Gupta A, Al-Kazwini H. Evaluating ChatGPT's diagnostic accuracy in detecting fundus images. Cureus, 2024,16(11):e73660.
[37] Pan XY, Bai Y, Zou K, et al. EH-Benchmark: Ophthalmic hallucination benchmark and agent-driven top-down traceable reasoning workflow. Inf Fusion, 2026,126:103631.
[38] Chen JS, Reddy AJ, Al-Sharif E, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? Ophthalmol Sci, 2025,5(1):100600.
[39] Kedia N, Sanjeev S, Ong J, et al. ChatGPT and Beyond: an overview of the growing field of large language models and their use in ophthalmology. Eye, 2024,38(7):1252-1261.
[40] Luo MJ, Pang JY, Bi SW, et al. Development and evaluation of a retrieval-augmented large language model framework for ophthalmology. JAMA Ophthalmol, 2024,142(9):798-805.
[41] Ke YH, Jin LY, Elangovan K, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med, 2025,8(1):187.
[42] Xiong ZX, Wang XF, Zhou YK, et al. How generalizable are foundation models whenapplied to different demographic groups and settings? Nejm Ai, 2025,2(1):AIcs2400497.
[43] Matta S, Lamard M, Zhang P, et al. A systematic review of generalization research in medical image classification. Comput Biol Med, 2024,183:109256.
[44] Hermann KL, Chen T, Kornblith S. The origins and prevalence of texture bias in convolutional neural networks. Adv Neural Inf Process Syst, 2020,33:19000-19015.
[45] AnaMaria Š. Measures of diagnostic accuracy: basic definitions. EJIFCC, 2009,19(4):203-211.
[46] 杨卫华, 邵毅, 许言午, 等. 眼科人工智能临床研究评价指南(2023). 国际眼科杂志, 2023,23(7):1064-1071.
[47] Gu J, Jiang X, Shi Z, et al. A survey on LLM-as-a-judge. The Innovation, 2026,7(6):101253.
[48] Porwal P, Pachade S, Kamble R, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data, 2018,3(3):25.
[49] Feng XR, Xu KZ, Luo MJ, et al. Latest developments of

generative artificial intelligence and applications in ophthalmology. Asia Pac J Ophthalmol (Phila), 2024,13(4):100090.
[50] Ramachandran R, Wey S. Mastering ophthalmology in the digital age. JAMA Ophthalmol, 2026,144(1):13-14.
[51] Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med, 2025,31(1):60-69.
[52] Betzler BK, Chen HC, Cheng CY, et al. Large language models and their impact in ophthalmology. Lancet Digit Heal, 2023,5(12):e917-e924.
[53] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med, 2022,28(5):924-933.

2025 版《中国科技期刊引证报告》核心版眼科学类期刊主要指标及排名

(以综合评价总分为序)

期刊名称	核心总被引频次		核心影响因子		综合评价总分	
	数值	排名	数值	排名	数值	排名
国际眼科杂志	**2278**	**1**	**1.067**	**2**	**75.3**	**1**
中华眼科杂志	1881	2	0.961	3	73.4	2
眼科新进展	1157	4	0.947	4	72.9	3
中国中医眼科杂志	1314	3	1.114	1	50.2	4
中华实验眼科杂志	877	5	0.593	8	49.3	5
中国眼耳鼻喉科杂志	446	8	0.624	6	48.4	6
中华眼底病杂志	609	7	0.603	7	46.1	7
中华眼视光学与视觉科学杂志	767	6	0.752	5	42.2	8
临床眼科杂志	329	9	0.359	10	36.3	9
中华眼科医学杂志电子版	151	12	0.113	12	32.7	10
中国斜视与小儿眼科杂志	243	11	0.493	9	27.5	11
眼科	301	10	0.237	11	22.6	12

摘编自 2025 版《中国科技期刊引证报告》核心版