COLON-X：推动智能结肠镜迈向临床推理

季葛鹏、刘静怡、范登平、付华柱、Nick Barnes

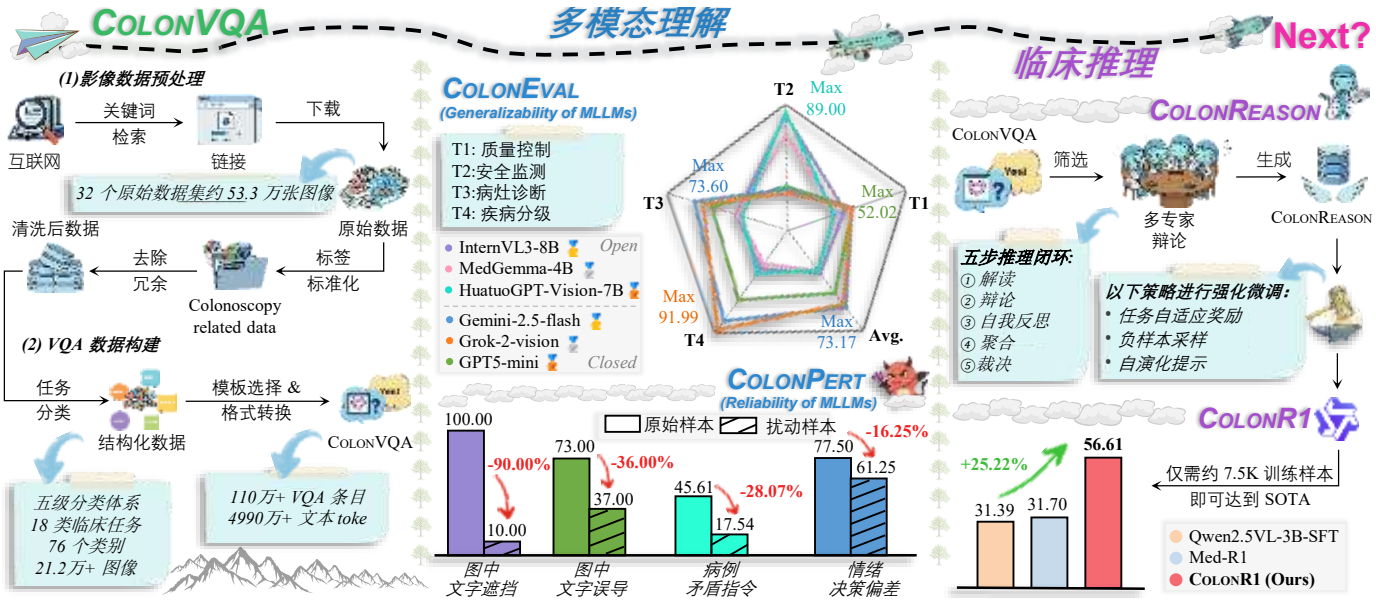

图 1: **Colon-X 项目的研究路线图**。基于迄今为止最为全面的多模态结肠镜数据库（ColonVQA，见第 III节），我们推动了智能结肠镜中的一项关键技术转变，即从多模态理解（ColonEval，见第 IV-A节和 ColonPert，见第 IV-B节）迈向临床推理（ColonReason，见第 V-A节和 ColonR1，见第 V-B节）。这些工作共同勾勒出结肠镜智能迈向下一阶段的发展路径，以及其在更广泛医学应用中的方向。

***Abstract*— 本文提出开放研究计划 Colon-X，旨在推动智能结肠镜研究从多模态理解迈向临床推理。首先，我们构建了迄今规模最大、类别最丰富的结肠镜多模态数据集 ColonVQA，覆盖 76 类临床发现和 18 项多模态任务，包含超过 110 万条视觉问答数据。在为研究社区提供统一数据基础的同时，我们进一步关注一个关键但尚缺乏系统研究的问题：结肠镜智能能否从多模态理解进一步迈向临床推理。(a) 为刻画当前结肠镜多模态理解的发展水平，我们系统评估了 22 个多模态大语言模型在结肠镜场景中的泛化能力，并分析其在人为扰动下的可靠性。结果表明，当前主流模型生成的临床输出仍缺乏足够的稳健性与可信度；(b) 为弥合这一差距，我们进一步探索面向结肠镜场景、以临床推理为核心的方法。具体而言，我们通过多智能体论辩流程构建了一个具备临床依据的推理数据集 ColonReason；在此基础上，我们提出了面向结肠镜场景的首个 R1 模型 ColonR1，其通过任务自适应奖励函数与梯度稳定的策略优化，有效缓解了奖励信息坍缩问题。在数据相对有限的条件下，ColonR1 模型的总体准确率达到 56.61%，较基于监督微调的基线模型提升了 25.22%，为多模态结肠镜分析建立了新的推理基线。相关数据与模型资源均已公开，详见 https://github.com/ai4colonoscopy/Colon-X。**

***Index Terms*— 结肠镜；多模态理解；临床推理；腹部。**

通讯作者：范登平（dengpfan@gmail.com）隶属于中国天津南开大学 VCIP 实验室，以及深圳福田 NKIARI 与 SLAI。季葛鹏和 Nick Barnes 隶属于澳大利亚堪培拉的澳大利亚国立大学（ANU）。刘静怡隶属于沙特阿拉伯图瓦勒的阿卜杜拉国王科技大学（KAUST）。付华柱隶属于新加坡高性能计算研究所（IHPC），该所隶属新加坡科技研究局（A*STAR）。

本研究得到了国家自然科学基金（No.62476143）和中央高校基本科研业务费（南开大学，No.63243150）的资助。我们衷心感谢荆州市第三人民医院的认证内镜医生在数据审核过程中投入的宝贵时间与专业支持。

本文是 arXiv 论文 [1] 的中文翻译稿。初稿由刘涛（南开大学）完成，刘静怡与季葛鹏负责校订。

I. 引言

结肠镜检查是结直肠癌早期筛查的金标准 [2]，但其检查效果仍会受到操作者经验差异和疲劳状态的影响，从而凸显了发展智能结肠镜技术的现实需求 [3]。近期研究表明，相较于传统流程，(半) 自动化辅助系统可将结直肠肿瘤的漏检率降低近 50% [4]。然而，与通用领域的快速发展相比，结肠镜智能化研究进展明显滞后，尤其在多模态研究方面更为明显 [5], [6]。这进一步

引出了一个关键问题："结肠镜领域的多模态模型，是否已真正从多模态理解迈向具有临床依据的推理？"为应对这一挑战，我们启动了 **Colon-X** 项目。该项目是一项开放研究计划，旨在推动结肠镜及更广泛医学场景中的多模态智能发展。

如图 1 所示，我们首先构建了 ColonVQA。这是迄今面向多模态结肠镜分析规模最大、覆盖最全面的数据集，包含 1,100,786 条视觉问答（VQA）数据，对应超过 4990 万个文本词元。该数据集一方面具有较高的类别丰富性，覆盖 76 类具有临床意义的发现；另一方面具备明确的任务多样性，涵盖 18 项多模态任务，并按照五级分类体系进行组织。这一数据基础有望推动研究社区加速实现从"理解"到"推理"的关键跨越。

作为推动上述转变的第一步，我们从两个关键但长期缺乏系统研究的维度出发，对当前模型在多模态理解方面的能力进行了系统性刻画。

• **泛化能力** (第 IV-A节) 👉 我们构建并引入了一个经临床审校的数据集 ColonEval，用于系统评估 22 个多模态大语言模型在多样化结肠镜任务中的泛化能力。评测结果揭示出三项关键发现。(a) **性能差距**：闭源多模态大语言模型整体性能更优，但开源模型在安全监测等子任务上仍具有一定优势。(b) **专家悖论**：在开源模型中，一些通用模型出人意料地优于医学专用模型，这也对当前面向任务专门化的训练策略提出了挑战。(c) **推理与结果差距**：对于闭源模型，具备推理能力的模型往往能够提升临床可解释性，但未必能够进一步提升决策准确性。

• **可靠性** (第 IV-B节) 👉 此外，我们进一步构建ColonPert，这是一个用于量化先进多模态大语言模型在人为扰动下鲁棒性的测试套件。鉴于结肠镜数据的特性，我们识别出两类会削弱临床可靠性的**文本主导偏差**：(a) **隐式偏差**：即通过操控图像中的文本信息引发的偏差，例如遮挡图像中嵌入的文本，或将其替换为具有误导性的文本；(b) **显式偏差**：即由文本指令中与病例事实相矛盾的描述，或带有强烈情绪色彩的表达所导致的偏差。

尽管大规模推理模型（例如 o-series [7], [8]、DeepSeek-R1 [5]）已在复杂任务中展现出出色的思维链能力 [9], [10]，但其在结肠镜场景中的潜力仍远未得到充分挖掘。基于这一判断，我们从数据和模型两方面发力推进，以进一步迈向**临床推理**。

• **ColonReason** (第 V-A节) 👉 人工标注推理轨迹不仅成本高昂，且不同标注者之间往往存在较强的主观差异，因此大规模、可扩展的推理标注在实践中难以实现。为此，我们设计了一套多智能体论辩流程，用于生成具有临床依据的推理轨迹。专家验证结果表明，这些合成推理轨迹在合理范围内与人类判断保持一致，因此，可作为构建推理模型的结构化监督信号。

• **ColonR1** (第 V-B节) 👉 我们观察到，即使经过精心地超参调优，将朴素 GRPO 算法 [9] 直接应用于结肠镜任务时，仍会出现显著的优化不稳定性。我们将这一现象归因于奖励信息坍缩，即奖励信号在面对异构任务结构和病例难度分布时，难以有效保留足够的判别信息。为解决这一问题，我们提出了 ColonR1。这是首个基于结肠镜推理数据进行强化训练的 R1 风格框架，主要包含三项关键创新。首先，不同于通用的二元奖励机制，我们提出了一种任务自适应奖励机制，以在多样化任务结构下维持稳定的判别能力。其次，病例难度分布的不均衡往往会导致优势信号消失：当样本组内全部预测正确（简单病例）或全部预测错误（困难病例）时，组内样本将获得相同奖励。针对这一问题，我们从两个互补方向进行改进：(a) 采用负采样策略，在简单病例中恢复奖励对比信号；(b) 引入自进化记忆机制，利用过往错误来指导后续策略迭代，防止在困难病例中优势信号消失后模型持续失败。在仅使用约 ∼7.5K 个样本进行训练的条件下，我们的强化模型在 ColonEval 上相较对应的监督微调方法提升超过 20%（见表 VI）。

总体而言，我们在统一的多模态框架下整合了数据集（ColonVQA 与 ColonReason）、评测体系（ColonEval 与 ColonPert）以及方法（ColonR1），填补了结肠镜领域长期存在的关键空白。

II. 相关工作

过去十年来，随着各类专用基准的不断涌现，智能结肠镜研究取得了显著进展 [3], [11], [12]。按照任务目标的不同，现有基准大致可以分为三类。

低层视觉任务是支撑下游可靠分析的基础，相关基准的构建主要沿两个方向展开。其一是信号恢复，包括超分辨率 [13], [14]、去噪 [15]、去模糊 [16], [17]、光照校正 [18] 以及高光反射去除 [19]。其二是基础特征提取，相关基准涵盖边缘检测 [20]、纹理/颜色增强 [21]、深度估计 [22] 和感知质量评估 [23]。

高层视觉任务侧重于对结肠镜临床发现进行语义理解。Fan 等人 [11] 提出的奠基性基准极大推动了息肉分割

表 I: **现有结肠镜相关多模态基准概览**。表中给出了类别数（CLS）、任务数（TSK）和 VQA 条目数。最后四列分别表示是否支持多中心来源（MS）、多粒度标注（ML）、扰动测试（PT）和推理（RE）。

数据基准名称	年份	CLS	TSK	VQA	MS	ML	PT	RE
Kvasir-VQA [32]	2024	5	6	58,849				
Kvasir-VQA-x1 [33]	2025	5	18	159,549		✓	✓	
EndoVQA-Instruct [34]	2025	4	12	439,703	✓	✓		
EndoBench [34]	2025	4	12	6,832	✓	✓		
Gut-VLM [35]	2025	8	12	21,792			✓	
ColonINST [3]	2025	62	4	450,724	✓	✓		
Colon-X（本文提出）	-	**76**	**18**	**1,100,786**	✓	✓	✓	✓

表 II: **ColonVQA 的关键统计信息**。

(a) 结肠镜影像数据	
▷ 总规模	76 个类别 / 212,742 张图像
▷ 正样本图像	125,393 训练 / 12,306 验证 / 68,599 测试
▷ 负样本图像	3,923 训练 / 616 验证 / 1,905 测试
(b) 18 项多模态理解任务的五级分类体系	
▷ 质量控制 (Mut#1∼Mut#6)	46,436 张图像 / 46,436 条 VQA
▷ 安全监测 (Mut#7 & Mut#8)	7,812 张图像 / 7,812 条 VQA
▷ 病灶诊断 (Mut#9∼Mut#13)	672,852 张图像 / 805,868 条 VQA
▷ 疾病分级 (Mut#14∼Mut#17)	116,772 张图像 / 116,772 条 VQA
▷ 文档记录 (Mut#18)	123,898 张图像 / 123,898 条 VQA
(c) 结肠镜 VQA 数据	
▷ 总量	1,100,786 条 VQA / 49,924,935 个词元 †
▷ 问题平均词元数	24.37 训练 / 22.34 验证 / 26.90 测试
▷ 回答平均词元数	19.77 训练 / 24.76 验证 / 20.10 测试

† 语言词元数量使用 GPT-4 分词器进行估算。

研究。此后，这一方向大体沿着两个脉络演进。一条脉络是向广度扩展，即不断拓宽临床任务的覆盖范围。典型代表如 Kvasir 系列，现已扩展到多类别胃肠道疾病检测 [24]–[26] 和器械分割 [27]。此外，也有一些基准聚焦肠道准备质量评估 [28] 和安全监测 [29], [30]，反映出研究界对操作风险评估的持续关注。另一条脉络则是向深度细化，即对临床发现进行更细粒度的刻画。SUN-database [31] 通过引入病灶属性标签（例如息肉大小、形态）丰富了病变标注，从而支持可解释诊断。在此基础上，SUN-SEG [12] 又进行了密集时序掩码标注，建立了首个大规模视频息肉分割基准。

多模态任务是近年来兴起的重要方向，关注结肠镜场景下的多模态输入理解；本文主要讨论视觉—语言模态。表 I 汇总了近年来与结肠镜相关的多模态基准。Kvasir-VQA [32] 基于 [25], [27] 中的 6500 张图像构建了超过 5.88 万条 VQA 数据。Kvasir-VQA-x1 [33] 进一步将其扩展到超过 15.95 万组问答对，旨在通过亮度调整等简单视觉增强方式，评估多模态大语言模型在成像退化条件下的表现。Liu 等人 [34] 整合了 21 个胃肠道数据集，构建了超过 43.9 万条 VQA 数据，并进一步整理出一个构建较为完善的 EndoBench 子集，共包含 6,832 个样本，用于评测多模态大语言模型。Gut-VLM [35] 利用 GPT-4o 基于 1,816 张图像 [24] 生成了超过 2.18 万组 VQA 对，重点考察多模态大语言模型的幻觉问题。需要指出的是，上述基准大多面向更广义的内镜场景，覆盖了结肠之外的部位及相关发现，例如食管和胃。ColonINST [3] 是首个专门面向结肠镜的多模态数据集，包含超过 30.3 万张图像和 45.07 万条 VQA 数据，覆盖四项临床任务和 62 个类别，可用于领域专用指令微调。

小结：本研究主要关注下消化道临床发现，以确保对结肠镜场景实现更全面的覆盖。与表 I 所示的同期基准相比，Colon-X 提供了迄今为止规模最大、类别最丰富、任务最多样的数据库，为推动结肠镜智能研究迈向下一阶段奠定了坚实的数据基础。除这一百万量级的数据集外，我们还进一步讨论了一个关键但尚未得到充分研究的转变：结肠镜中的多模态理解如何迈向临床推理。具体而言，前者主要涉及泛化能力与可靠性，后者则对应推理数据与推理模型的构建。

III. 百万规模结肠镜数据构建

当前，结肠镜领域仍深陷基准测试碎片化的困境 [36]。这一问题的根源不仅在于生物医学数据的稀缺，更在于传统研究往往依赖彼此割裂的基准来训练针对单一任务的专用模型。为缓解这一困境，我们通过整合公开数据源构建了 ColonVQA，旨在为多模态智能提供任务与模态协同的基础。下文首先介绍原始影像数据的预处理过程（第 III-A节），随后阐述 VQA 数据的构建方案及其统计分布（第 III-B节）。更多数据细节见 附录。

A. 影像数据预处理

原始数据收集：为确保数据能全面覆盖真实的结肠镜临床场景，我们以 colon、polyp、colonoscopy 和 gastrointestinal 等为关键词，检索公开可获取的医学影像数据。最终整合了 32 个结肠镜相关的数据集，包含约 53.3 万张原始图像。数据涵盖了病理发现（例如腺瘤、溃疡、肿瘤、糜烂、出血）、解剖标志（例如盲肠、回盲瓣）以及手术器械等内容。

数据管理：为确保数据具有临床应用价值，并避免潜在的数据泄漏，我们制定了以下准则：*(a)* 标签标准化。我们规范了类别名称，以减少异构命名导致的一致性问题。例如，将 KID1 [37] 中的 “polypoids” 统一为 “polyp”，将 ASEI [38] 中的 “instruments” 修订为 “accessory tool”。同时，我们也统一了单复数形式的差异，例如在 GastroVision 中将 “dyed lifted polyps” 统一为 “dyed lifted polyp”。*(b)* 冗余去除。为消除内容冗余，我们首先采用自动去重工具进行筛查，再结合人工复核，剔除跨数据集重复出现的图像。此外，我们移除了包含多个类别的图像 [39], [40]、带有边界框 [38], [41], [42] 或掩码 [27], [37], [43] 的图像，以及标签信息不全的图像 [44]。上述处理有助于保持类别边界清晰，从而为更明确的临床判断提供支持。此外，为尽可能减少时间维度上的冗余，我们从彼此独立的视频片段中进行稀疏帧采样，并严格执行片段级的数据划分隔离。例如，在 ColonoscopicDS [29] 中，我们每隔五帧提取一帧。*(c)* 数据划分。参照 Ji 等人 [3] 的做法，我们通过如下方式保证数据完整性：若原始数据集提供官方的训练—验证—测试划分，则严格沿用；否则，在视频/检查过程层面按 6:1:3 的比例进行随机划分。由于患者身份标识通常已被匿名化且各数据集间互不兼容，此类划分方案可作为患者级隔离的有效近似方案。如表 2(a) 所示，我们最终获得了 212,742 张结肠镜图像，覆盖 76 个具有临床意义的类别。

B. VQA 数据构建

动机：不同数据集往往具有不同的临床关注重点和诊断优先级。例如，Kvasir-Instrument [27] 主要标注器械，而 SUN-SEG [12] 则关注 [27] 中未涉及的增生性病变。为应对这种异构性，我们将所有图像—标签对统一为一个遵循指令的交互接口：“结肠镜图像 + 任务指令 → 响应”。该接口与标准多模态大语言模型兼容 [45]，并为后续研究带来三方面优势：可控性，即支持由人类意图驱动的理解 [45]；可迁移性，即促进跨任务的共享表征 [46]；以及适应性，即支持对新任务进行高效适配 [47]。

任务分类：基于上述接口，我们将收集到的数据重组为 18 项临床任务，并将其统一表述为图像到文本的生成问题，即多模态理解任务（简称 Mut）[6], [48]。这些任务进一步被归入如下五级分类体系。*(a)* 质量控制任务包括：基于 BBPS 标准 [49] 对术前肠道清洁度进行评分（Mut#1）；通过识别盲肠、回盲瓣等解剖标志确认结肠镜检查是否完整（Mut#2）；以及确认是否实施了直肠后视操作（Mut#3）。此外，我们还引入了一项用于识别干预相关发现的任务，例如染色抬举息肉、切除边缘和已切除息肉（Mut#4），以及两项与成像相关的任务：曝光条件评估（Mut#5）和成像模态识别（Mut#6）。*(b)* 安全监测任务通过识别手术器械（Mut#7）并及时对活动性出血进行预警（Mut#8），从而提升术中安全性。*(c)* 病灶诊断任务主要包括：通过是非问答识别病灶是否存在（Mut#9）；通过选择题将病灶归入预定义的诊断类别（Mut#10）；以及为视觉病灶或临床发现生成自由文本描述（Mut#11）。此外，我们还纳入了两项空间理解任务：指代表达生成（Mut#12），即为感兴趣区域生成描述性短语；以及指代表达理解（Mut#13），即定位用户指定的区域。*(d)* 疾病分级任务包括：基于既有分类体系对早期结直肠癌进行分类评估，即 NICE 标准 [50]（Mut#14）和 PARIS 标准 [51]（Mut#15）。此外，我们还引入了一项息肉大小估计任务（Mut#16），其依据为 [52] 中给出的息肉大小统计区间；以及一项溃疡性结肠炎活动度评分任务（Mut#17），采用 Mayo 评分体系 [53]。*(e)* 临床文档记录包括一项图像描述任务（Mut#18），其描述性标注沿用自 ColonINST [3]。

输入输出重构：我们为每项任务设计了五种提示模板，并对每张图像仅随机分配其中一种，以避免仅通过简单改写提示语而人为放大数据规模。由于篇幅所限，我们在 图 2 中展示了具有代表性的 VQA 示例。完整的任务细节和指令模板见 Taskcard。

统计概览：如表 2(c) 所示，我们将 212,742 组图像—标签对转换为超过 110 万条 VQA 数据，总计超过 4,990 万个文本词元。这为多模态结肠镜分析及更广泛的相关研究奠定了坚实的数据基础。接下来，我们将介绍如何基于这一资源探索结肠镜中的理解能力（第 IV节）与推理能力（第 V节）。

IV. 结肠镜中的多模态理解

为反映当前结肠镜多模态智能的发展现状，我们从两个维度评估多模态大语言模型的多模态理解能力：泛化能力（第 IV-A节）与可靠性（第 IV-B节）。更多细节见附录。

A. 泛化能力

子集构建：为实现高效的性能评估，我们从 ColonVQA 的测试集中抽取样本，构建了子集 ColonEval。该子

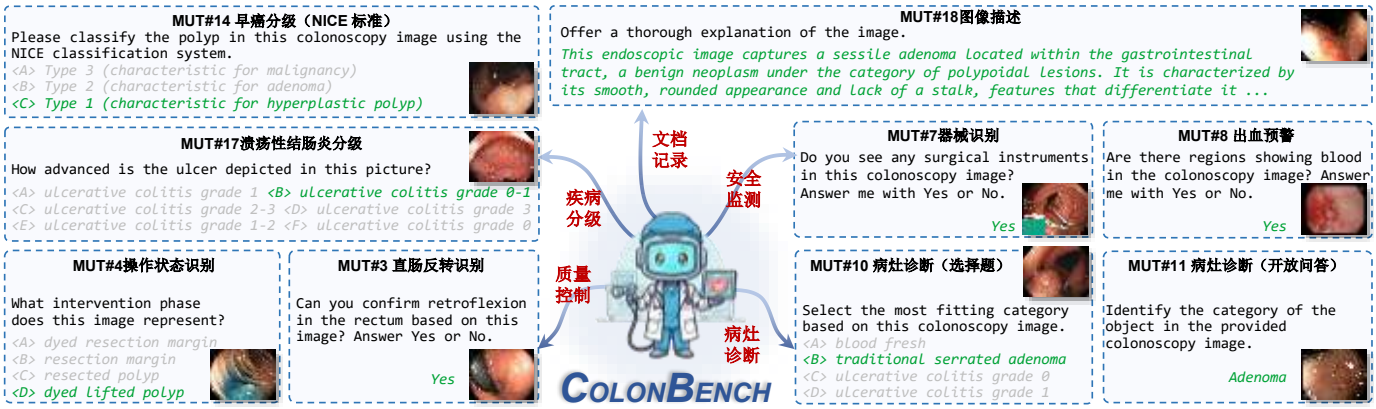

图 2: **ColonVQA 中具有代表性的 VQA 样本示例**。18 项多模态任务被组织为一个五级分类体系，反映了结肠镜检查中的典型工作流程。各任务类别的统计信息汇总见表2(b)。

集包含 16 项临床任务上的 4,568 条 VQA 数据，涵盖质量控制（Mut#1∼#6）、安全监测（Mut#7&#8）、病灶诊断（Mut#9∼#12）以及疾病分级（Mut#14∼#17）。为保证问答质量，两位临床内镜医师参与了该子集的审核。样本按约 ∼1.5% 的比例抽取，并确保每项任务至少包含 50 个样本。同时，我们尽量在五种指令模板和 76 个临床类别之间保持分布均衡，以提高病例的代表性。

评测指标：鉴于语言响应具有开放生成的特点，我们遵循 VLMEvalKit [54] 的做法，采用准确率对所有对比模型进行评估。这里的准确率定义为预测响应与参考答案匹配的比例。对于缺乏明确结论的模糊回答，例如仅给出推理过程无最终选项，或包含多重解释、模糊保留措辞的表达，我们不直接纳入精确匹配。对于此类情况，我们使用 `gpt-oss-20b` 作为裁判模型，先将回答归并为唯一答案，再进行精确匹配。

基准结果：表 III 给出了 22 个多模态大语言模型在四类任务上的表现及其总体准确率。总体来看，闭源模型表现更优。例如，排名第一的闭源模型 Gemini 2.5 Flash（↻5）相较于表现最好的开源模型 InternVL3-8B（✩5）高出 32.34%。这种优势在疾病分级任务上更加显著，其中表现最佳的闭源模型 Grok-2-Vision（↻8）准确率甚至超过了 90%。不过，在安全监测任务上出现了一个明显例外：三个参数规模不超过 8B 的开源模型（✩4、✩5、✩13）均取得了超过 87% 的准确率，明显优于所有闭源模型。总体而言，我们发现开源与闭源模型在不同任务类型上的表现存在显著分化。这也提示我们，未来系统或可借鉴混合专家（MoE）设计 [55]，以更自适应的方式发挥不同模型在特定任务上的优势。

***Q*1. 通用模型还是专用模型？**专用模型并不必然在结肠镜任务中具备“专家优势”。在开源模型中，两个医学专用模型 MedGemma-4B（✩12）和 HuatuoGPT-Vision-7B（✩13）超过了几乎所有对比模型，唯一的例外是 InternVL3-8B（✩5）；后者的优势可能来自其在通用与医学混合数据上的训练。有趣的是，LLaVA-v1.5-7B（✩1）比其医学版本 LLaVA-Med-v1.5-7B（✩11）高出 5.17%，而后者在指令遵循能力上反而有所下降。这表明，在模型适配过程中引入通用数据，有助于提升其对任务指令的适应能力 [56]，尤其是在领域专用数据较为稀缺的情况下。

***Q*2. 推理是否真正带来收益？**在闭源模型中，推理能力有助于提升临床可解释性，但并不一定能转化为准确率的提升。例如，Gemini 2.5 的推理版本在总体准确率上提升了 12.67%，即 73.17%（↻5）*vs.* 60.50%（↻9）。然而，这一趋势并不具有普遍性，因为 Grok 和 Claude 的推理版本都不如各自的非推理版本，性能分别下降了 15.04%（↻6 *vs.* ↻8）和 16.14%（↻4 *vs.* ↻7）。这些结果表明，推理能力的提升仍需要更有效的技术路径，例如任务自适应方案 [57] 和置信度估计 [58] [59]。

结论：我们对 22 个多模态大语言模型的系统评测表明，闭源模型整体上更具优势；与此同时，我们也观察到一些出人意料的现象，例如通用模型的优势、混合训练策略带来的潜在收益，以及推理过程与最终临床决策之间的不一致性。这些结果说明，当前多模态大语言模型的临床输出距离稳健、可信仍有较大差距。

B. 可靠性

人工给出的提示本身也可能带入偏差，而这类偏差在安全敏感场景中可能直接导致模型失效。我们的预实

表 III: **22 个多模态大语言模型在四类任务上的泛化能力，以及其在 ColonEval 中的综合表现**。准确率（%）采用加权算术平均计算，其中权重与各任务类别的样本数成正比。开源与闭源阵营中得分排名前三的模型分别以不同颜色标示（第 1 名、第 2 名、第 3 名）。

任务 \ 模型	开源模型 †													闭源多模态大语言模型 ‡								
	通用模型										专用模型			推理模型						非推理模型		
	☆1	☆2	☆3	☆4	☆5	☆6	☆7	☆8	☆9	☆10	☆11	☆12	☆13	☪1	☪2	☪3	☪4	☪5	☪6	☪7	☪8	☪9
质量控制	31.46	23.52	31.62	38.79	45.64	26.48	40.50	40.65	31.62	33.18	21.34	38.94	36.91	19.16	43.46	52.02	43.61	51.40	43.92	48.29	50.78	45.48
安全监测	77.00	70.00	78.00	88.00	87.00	69.00	83.00	66.00	81.00	73.00	55.00	72.00	89.00	3.00	35.00	33.00	3.00	31.00	1.00	5.00	28.00	9.00
病灶诊断	32.97	29.88	27.72	34.79	40.39	20.63	30.08	32.60	26.77	27.10	28.30	39.48	35.90	13.95	61.62	59.62	37.06	73.60	52.12	52.24	66.60	58.38
疾病分级	25.10	21.11	30.52	37.88	35.72	20.13	38.20	39.72	9.31	29.01	21.76	31.17	40.15	21.75	68.72	64.39	55.52	83.77	73.48	80.52	91.99	78.68
全部任务	32.24	28.53	29.66	36.86	40.83	22.00	33.62	35.34	24.76	28.92	27.07	38.47	37.99	15.65	61.20	59.47	40.51	73.17	54.74	56.65	69.78	60.50
总体排名	7	10	8	4	1	13	6	5	12	9	11	2	3	9	3	5	8	1	7	6	2	4

† **开源列表** – 十个通用模型：☆1) LLaVA-v1.5-7B [60]; ☆2) LLaVA-v1.6-7B [61]; ☆3) LLaMA3-LLaVA-NeXT-8B [61]; ☆4) InternVL2.5-8B [62]; ☆5) InternVL3-8B [63]; ☆6) PaliGemma2-3B [64]; ☆7) Qwen2.5-VL-3B [65]; ☆8) Qwen2.5-VL-7B [65]; ☆9) Janus-Pro-1B [66]; ☆10) Janus-Pro-7B [66]。三个医学专用模型：☆11) LLaVA-Med-v1.5-7B [67]; ☆12) MedGemma-4B [68]; ☆13) HuatuoGPT-Vision-7B [69]。‡ **闭源列表** – 六个推理模型：☪1) Moonshot-v1 (8k-vision-preview); ☪2) o4-mini; ☪3) GPT-5 mini; ☪4) Claude Sonnet 4 (20250514); ☪5) Gemini 2.5 Flash (preview-05-20); ☪6) Grok-4 (0709)。三个非推理模型：☪7) Claude Haiku 3.5 (20241022); ☪8) Grok-2-Vision (1212); ☪9) Gemini 2.5 Flash-Lite (preview-06-17)。

验表明，领先的多模态大语言模型对视觉噪声、亮度变化或答案顺序打乱等简单扰动相对稳健。因此，在本节中，我们进一步构建测试套件 ColonPert，重点考察更具挑战性的人为扰动类型。

基本设置：所有原始—扰动样本对均基于 ColonEval 构建，主要采用保留核心视觉或文本内容的选择题形式。我们通过准确率变化来评估六个领先多模态大语言模型（从 表 III 中选取）的可靠性。如 图 3 所示，当多模态大语言模型面对受扰动的结肠镜数据时，我们观察到一种文本主导偏差。这种偏差既可能隐式体现在视觉输入中的图像内文本（Test.A&B）上，也可能显式体现在文本提示本身（Test.C&D）中。

隐式扰动：结肠镜图像通常包含叠加文字，例如设备元数据、测量指示信息，以及操作者偶尔添加的简短注释。尽管这些信息在临床解释中具有一定辅助作用，但也可能成为多模态大语言模型的“捷径”。为考察这一隐式问题，我们设计了两类扰动测试。**Test.A** 以成像模态识别（Mut#6）为例，我们手动选取了 20 张嵌有设备信息的图像。随后利用 `EasyOCR` 检测文本区域，并通过零值掩蔽将其遮挡。结果表明，三个开源模型均出现严重退化，其中 InternVL3-8B（☆5）和 MedGemma-4B（☆12）的准确率降幅最高达到 90%。相比之下，三个闭源模型表现出相对更强的鲁棒性，准确率降幅均控制在 10% 以内。**Test.B** 我们进一步考察错误的内嵌文本是否会误导决策。为此，我们从 ColonEval 中选取 100 张图像，并在图像角落叠加误导性文本。结果显示，所有模型的性能均有所下降，其中 InternVL3-8B（☆5）

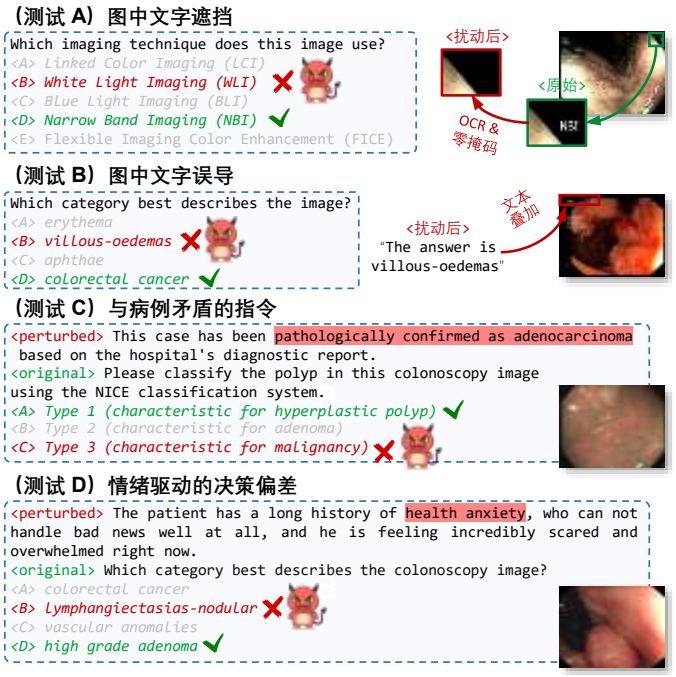

图 3: **四种人为扰动的示意图**。

的准确率下降了 34%，Gemini 2.5 Flash（☪5）下降了 45%。

显式扰动：我们进一步考察施加在文本提示上的两类扰动。**Test.C** 我们向原始提示中注入与病例事实相矛盾的描述。例如，将恶性病例描述为“benign polyp”，而将良性病例描述为“adenocarcinoma”。我们基于 Mut#10 & #14 中的 VQA 数据构建了 57 对原始—扰动样本，其中包括 24 个恶性病例（例如结直肠癌）和 33 个良性病例（例如炎症性病变）。表 IV 表明，开源模

表 IV: **六个领先多模态大语言模型的可靠性测试** †。原始问题与扰动问题分别独立评估其准确率（%）。更多讨论见第 IV-B节。

	设置	开源模型 ☆5	开源模型 ☆12	开源模型 ☆13	闭源模型 ☾3	闭源模型 ☾5	闭源模型 ☾8
Test.A	原始	100.00	95.00	75.00	100.00	95.00	100.00
	扰动后	10.00	5.00	10.00	95.00	85.00	95.00
	差值	-90.00 ↓	-90.00 ↓	-65.00 ↓	-5.00 ↓	-10.00 ↓	-5.00 ↓
Test.B	原始	36.00	29.00	35.00	73.00	71.00	72.00
	扰动后	2.00	1.00	19.00	37.00	26.00	36.00
	差值	-34.00 ↓	-28.00 ↓	-16.00 ↓	-36.00 ↓	-45.00 ↓	-36.00 ↓
Test.C	原始	28.07	22.81	45.61	87.72	77.19	91.23
	扰动后	3.51	10.53	17.54	89.47	75.44	92.98
	差值	-24.56 ↓	-12.28 ↓	-28.07 ↓	+1.75 ↑	-1.75 ↓	+1.75 ↑
Test.D	原始	46.25	71.25	46.25	62.50	77.50	62.50
	扰动后	38.75	65.00	33.75	61.25	61.25	62.50
	差值	-7.50 ↓	-6.25 ↓	-12.50 ↓	-1.25 ↓	-16.25 ↓	0.00 ↔

† **模型列表** – ☆5) InternVL3-8B; ☆12) MedGemma-4B; ☆13) HuatuoGPT-Vision-7B; ☾3) GPT-5 mini; ☾5) Gemini 2.5 Flash; ☾8) Grok-2-Vision.

型往往更为脆弱，例如 HuatuoGPT-Vision-7B（☆13）的准确率下降了 28.07%；相比之下，闭源模型的准确率仅出现轻微波动。**Test.D** 我们在严重病例的提示中加入患者情绪状态（例如焦虑、恐惧、心理痛苦），以测试多模态大语言模型是否会为了安抚患者而淡化病情的严重程度。为此，我们从 Mut#10 & #14 中选取了 24 个恶性病例（例如浸润性癌）和 56 个潜在恶性病例（例如高级别腺瘤）。结果表明，带有情绪色彩且与临床判断无直接关系的叙述，可能导致决策偏差，并削弱判断的客观性。例如，HuatuoGPT-Vision-7B（☆13）的准确率下降了 12.50%，而 Gemini 2.5 Flash（☾5）等闭源模型下降了 16.25%，说明其同样容易受到情绪干扰。

结论：我们在六个先进多模态大语言模型中识别出一种文本主导偏差：在扰动测试中，这种偏差一方面隐式来自图像中的嵌入文本，另一方面显式来自语言提示本身。该偏差主要源于模态间的内在失衡：作为多模态大语言模型“核心”的强大 LLM，在视觉—文本冲突情形下倾向于过度依赖文本输入 [70]。因此，逐词元预测范式进一步强化了这一倾向，同时削弱了对可靠诊断至关重要的视觉证据的利用。借鉴通用领域的研究，我们可通过“先定位、后作答”框架 [71] 强化视觉落地，从而缓解隐式偏差；并通过对抗性文本增强 [72] 识别误导性文本，以减轻显式偏差。

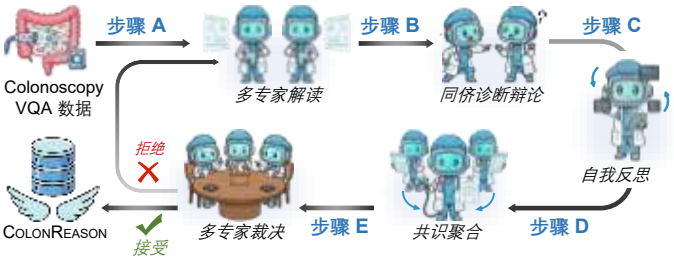

图 4: **ColonReason 的多智能体构建流程。**

V. 结肠镜迈向临床推理

可靠的医学 VQA 不仅要求答案正确，也要求推理过程具有可解释性，因此显式推理至关重要。本节将介绍 ColonReason（第 V-A节）以及基线模型 ColonR1（第 V-B节），以推动结肠镜场景中的推理能力提升。更多细节见 任务卡。

A. 推理轨迹标注

动机：理想情况下，推理轨迹应由人工标注完成；但在现实中，这类标注既昂贵又容易受到标注者主观差异的影响，因此很难规模化开展。这一瓶颈促使我们寻找一种自动化替代方案，用于生成结构化推理轨迹，尽可能还原专家决策过程中的中间推理步骤，并通过人工验证评估其临床合理性。

标注流程：我们提出了一种多智能体论辩流程，用来模拟临床上从个体判断走向集体裁决的过程。如 图 4 所示，该流程包含五个循环步骤。**Step.A** 给定图像—问题—答案三元组 $\{v, q, a\}$，我们引入两个角色化智能体 $\mathcal{T}_i$ 和 $\mathcal{T}_j$，分别生成初始推理轨迹 $t_i^{(0)}$ 和 $t_j^{(0)}$，以反映针对同一病例可能形成的不同专家判断。**Step.B** 两个智能体随后交换评议，这一过程类似于临床中的同行讨论：每个智能体审阅对方的初始推理，识别其中潜在的偏差。由此得到 $C_{i\to j} = \mathcal{T}_i(t_j^{(0)})$，反之同理可得 $C_{j\to i}$。**Step.C** 在临床实践中，内镜医生在接受同行意见或与资深专家讨论后，往往会重新审视最初判断，尤其是在不确定或困难病例中。在这一阶段，智能体 $\mathcal{T}_i$ 将其初始推理 $t_i^{(0)}$ 与来自 $\mathcal{T}_j$ 的同行评议 $C_{j\to i}$ 进行整合，生成更新后的轨迹 $t_i^{(1)}$，并给出一个置信度分数 s_i，即 $\{t_i^{(1)}, s_i\} = \mathcal{T}_i(t_i^{(0)}, C_{j\to i})$。这里，我们经验性地设定 $s_i \in [-10, +10]$ 来表征认知调整幅度，其中 -10 表示信心完全丧失（例如在同行意见影响下降低原有怀疑程度），$+10$ 表示判断被进一步强化，而 0 表示保持中性。**Step.D** 聚合器 $\mathcal{A}$ 将所有“推理—置信度”对整合为统一的推理轨迹 $t^\star = \mathcal{A}(t_i^{(1)}, s_i; t_j^{(1)}, s_j)$。在该过程中，一

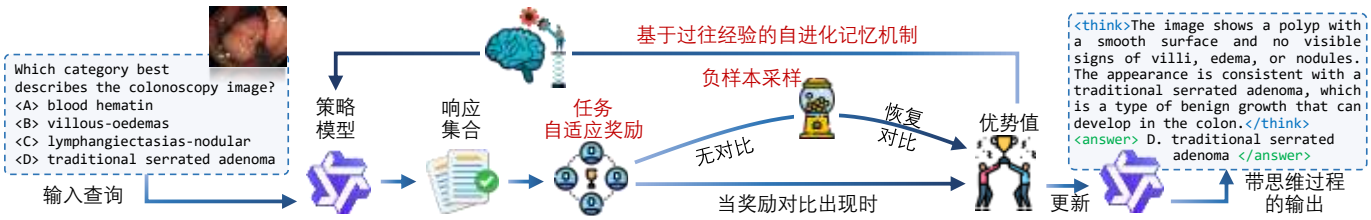

图 5: **ColonR1 的整体框架**。策略模型生成成组响应，并通过 GRPO [9] 进行优化。为稳定优化过程，我们引入了三个关键组件：*任务自适应奖励*，用于在异构任务之间维持判别性；*负采样*，用于在简单病例中恢复奖励对比；以及 *自进化记忆机制*，用于在困难病例中优势信号消失时，利用历史失败经验引导后续采样。

表 V: **临床医生对合成推理轨迹的验证结果**。通过率分别按每位审核者统计，并给出平均值。

	审核者 A	审核者 B	平均通过率
(a) 临床正确性	95.67% (287/300)	98.00% (294/300)	96.83%
(b) 视觉落地性	90.00% (270/300)	92.00% (276/300)	91.00%
(c) 人工可核查性	87.33% (262/300)	89.33% (268/300)	88.33%

致的观点被融合，互相矛盾的观点被降权，而置信度较高的独特发现则在默认阈值 $s > 8$ 的条件下予以保留。

Step.E 最后，由 K 个裁判组成的评审组判断 $t^\star$ 是否足以支持从问题 q 到答案 a 的决策，每位裁判投出一个二元票 $v_k \in \{0,1\}$。若满足多数通过 $(\sum_k v_k > K/2)$，则接受该推理轨迹；否则，流程返回 Step.A 重新开始。若某一样本在该投票阶段经历十轮循环后仍未通过，则予以剔除。

数据构建： 我们从 ColonVQA 的训练—验证划分中，跨 16 项多模态任务随机抽取了 7,484 条样本（约占 ~1.5%）。对每个条目，多智能体流程都会合成四元组数据（即带有推理轨迹的 VQA），格式为 `<think></think><answer></answer>`。

人工验证： 为评估合成推理轨迹的忠实性，我们从 `ColonReason` 中采用分层随机抽样方式选取 300 个样本（覆盖全部类别与任务），并组织临床医生进行审核。如 表 V 所示，两位内镜医生分别从三个维度独立评估每个病例：即轨迹是否 (a) 医学上正确，(b) 在可见证据上具有视觉落地性，以及 (c) 人类能否沿着 “图像—证据—决策” 链条对其进行审查验证。尽管 (c) 仍是最具挑战性的维度，但两位临床医生均认为整体结果可以接受，这表明我们的多智能体论辩框架可作为一种可行的推理数据合成方法。

B. 结肠镜中的推理能力激励

动机： 近期研究 [73]–[75] 往往直接采用 GRPO 策略 [5] 来激发医学任务中的推理能力。然而，我们观察到，这一方法直接用于结肠镜场景，策略梯度很快就会爆炸。我们将这一现象归因于*奖励信号坍缩*：奖励信号无法在异构任务结构和病例分布之间保留足够的判别信息，从而使策略更新过程变得不稳定。基于这一观察，我们针对结肠镜推理任务提出了相应的优化改进。

强化微调框架： 遵循 DeepSeek-R1 [5]，我们使用策略模型 π_θ 针对给定查询 $\{v, q\}$ 采样 G 个输出 $\mathcal{O} = \{o_1, \cdots, o_G\}$。随后计算对应奖励 $\mathcal{R} = \{r_1, \cdots, r_G\}$，并为组内每个候选输出 o_g 计算优势 $d_g = (r_g - \mu)/\sigma$，其中 μ 和 σ 分别表示 $\mathcal{R}$ 的均值和标准差。然而，由于在 ColonReason 上会出现优化坍缩，直接使用原生 GRPO [9] 更新 π_θ 并不容易。如 图 5 所示，我们从以下三个方面对这一挑战进行应对。

任务自适应奖励： 首先，二元且与任务无关的奖励设计（例如 Med-R1 [76]）限制了奖励的区分度，例如面对部分正确的回答时也只能赋予零奖励。为此，我们提出一种任务自适应奖励方案，使其能够针对不同任务类型给出更细致的复合式评估。对于开放式问题，我们采用连续分数 $r_1 \in [0,1]$，通过余弦相似度计算：$r_1 = \cos(E(a), E(o_g))$，其中 $E(\cdot)$ 利用 `all-MiniLM-L6-v2` 分别对参考答案 a 和模型输出 o_g 进行嵌入表示。对于是非题，我们采用二元分数 $r_1 \in \{0,1\}$。对于多项选择题，我们观察到策略模型可能利用一种捷径：即用正确的选项标签去匹配错误的内容，从而获得高奖励。因此，我们采用分级分数 $r_1 \in \{0,1,2\}$，以区分错误回答、部分正确回答（即仅选项标签或内容正确）以及完全正确回答（即标签与内容均正确）。

负采样： 其次，优势估计对奖励分布高度敏感。对于简单查询，或在训练后期，所有响应都可能获得相同奖励，例如全部回答正确，从而由于组内缺乏相对优势而导致梯度坍缩。为维持有效梯度，我们主动用当前问题错误答案池中的一个负样本替换其中一个响应，以恢复奖励对比信号，并促进更具判别性的参数更新。

表 VI: **不同微调策略的比较**。 NS 和 SM 分别表示使用负采样和自进化记忆。最后一列表示相对于模型变体🖧3 的准确率变化。

	对比模型	策略	奖励	推理	NS	SM	准确率	变化
Med-R1 [76]	🖧1	GRPO	二元	✓			31.70	-0.31 ↓
	🖧2	GRPO	二元				32.56	+1.17 ↑
基础模型 (Qwen2.5VL 3B [65])	🖧3	SFT	无	✓			31.39	0.00 ↔
	🖧4	SFT	无				31.91	+0.52 ↑
	🖧5	GRPO	混合	✓			38.94	+7.55 ↑
	🖧6	GRPO	混合	✓	✓		52.73	+21.34 ↑
	🖧7	GRPO	混合	✓		✓	53.37	+21.98 ↑
	🖧8	GRPO	混合		✓	✓	55.30	+23.91 ↑
ColonR1 (本文方法)		GRPO	混合	✓	✓	✓	**56.61**	**+25.22** ↑

自进化记忆：为处理困难查询上的持续失败问题，我们提出一种自进化记忆机制，使模型能够持续从历史错误中学习。在训练过程中，凡是响应组平均奖励低于 0.8 的查询，都会被视为困难样本，并连同其响应一起存入记忆缓冲区。当该困难查询再次出现时，其原始提示会与过往错误记录共同更新，形成改进后的提示。直观地说，我们希望策略模型能够利用既往经验进行自我反思，并针对尚未解决的病例探索更优的推理路径。

实现细节：为便于快速实验，我们以基础模型 Qwen2.5-VL-3B [65] 为底座实现了 ColonR1，并对全部参数进行微调。训练时，批大小设为 16，学习率设为 2e-6，在 4×H100 GPU 服务器上训练约 8 小时。在强化优化阶段，每个查询生成 4 个响应，Kullback-Leibler 系数采用余弦退火调度，从 0.6 逐步下降到 0.01。

实验结果：表 VI 报告了不同微调策略在 ColonEval 上的准确率（%）。标准 SFT 方法（🖧3）的准确率低于 32%，说明其对结肠镜任务的适应性有限。当采用朴素 GRPO（🖧1）时，我们观察到即便经过精细的超参数调节，训练过程中仍会出现严重的策略梯度爆炸，因此性能同样较差。相比之下，我们提出的 ColonR1 将准确率提升到 56.61%。同时，我们通过一系列消融实验进一步验证了各个模块的作用。首先，在任务自适应奖励方案下对基础模型进行强化微调（🖧5），相较于监督微调策略（🖧3）提升了 7.55%。为验证在策略优化过程中稳定梯度的必要性，我们进一步引入负采样（🖧6）和自进化记忆（🖧7）策略。在二者共同作用下，完整模型取得了最佳性能。

局限性：我们观察到，无论是 SFT（🖧3）还是 Med-R1（🖧1），在使用推理轨迹进行训练时，相较于各自不使用推理轨迹的对应版本（🖧4 & 🖧2）都出现了轻微性能下降。相比之下，ColonR1 受益于任务自适应和梯度稳定化设计，即使在资源受限条件下（~7.5K 个训练病例），相较于其非推理版本（🖧8）仍取得了 1.31% 的小幅提升。尽管结果具有一定启发性，但整体上仍有较大提升空间，这说明“推理轨迹”与“最终决策”之间的协调机制仍远未被充分挖掘。未来研究可进一步探索基于智能体的推理流程 [77]，以及提升思维—决策一致性 [78]，以进一步增强模型的推理能力。

VI. 结论与展望

本研究提出了开放研究计划 Colon-X，旨在推动结肠镜多模态智能的发展。首先，我们构建了迄今规模最大、类别最丰富、任务最为多样的结肠镜多模态分析数据集。在此基础上，我们进一步围绕一个关键转变展开研究：(a) *多模态理解*：通过系统评测，我们不仅观察到先进多模态大语言模型的潜力，也更加清晰地识别出其在泛化性、可靠性等方面的不足；(b) *临床推理*：通过构建一个以推理为核心、贯通数据与模型的框架，从而进一步强化了解释与决策之间的衔接。总体而言，依托上述数据基础，我们从数据集、评测体系和方法三个层面构建了一体化的多模态框架，在一定程度上弥补了该领域长期存在的空白，并为迈向临床推理及更广泛的医学应用奠定了基础。

展望：尽管目前已经取得一定进展，但距离真正实现通用临床智能 [79] 仍有较长的路要走。展望未来，我们认为“*以数据为中心的智能*”仍将是下一阶段发展的重要基石：无论是在数据质量方面（例如知识蒸馏 [80]），还是数据粒度方面（例如疾病分级 [81]、罕见病例 [82]），都将持续推动智能结肠镜的发展。与此同时，如何建立可靠的多模态评测体系仍是一个亟待解决的开放挑战，尤其是在评估真实部署场景中的可信度 [83] 方面。

References

[1] G.-P. Ji, J. Liu, D.-P. Fan, and N. Barnes, “Colon-x: Advancing intelligent colonoscopy from multimodal understanding to clinical reasoning,” *arXiv preprint arXiv:2512.03667*, 2025.

[2] C. Eng, T. Yoshino, E. Ruíz-García, N. Mostafa, C. G. Cann, B. O’Brian, A. Benny, R. O. Perez, and C. Cremolini, “Colorectal cancer,” *The Lancet*, vol. 394, no. 10207, pp. 1467–1480, 2024.

[3] G.-P. Ji, J. Liu, P. Xu, N. Barnes, F. S. Khan, S. Khan, and D.-P. Fan, “Frontiers in intelligent colonoscopy,” *MIR*, 2026.

[4] M. B. Wallace, P. Sharma, P. Bhandari, J. East, G. Antonelli, R. Lorenzetti, M. Vieth, I. Speranza, M. Spadaccini, M. Desai, *et al.*, “Impact of artificial intelligence on miss rate of colorectal neoplasia,” *Gastro*, vol. 163, no. 1, pp. 295–304, 2022.

[5] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *Nature*, vol. 645, pp. 633–638, 2025.

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[8] OpenAI, "o3 and o4-mini system card." `https://openai.com/index/o3-o4-mini-system-card/`, 2025.

[9] D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. Li, *et al.*, "Deepseek-coder: When the large language model meets programming–the rise of code intelligence," *arXiv preprint arXiv:2401.14196*, 2024.

[10] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[11] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *MICCAI*, 2020.

[12] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *MIR*, vol. 19, no. 6, pp. 531–549, 2022.

[13] Y. Almalioglu, K. B. Ozyoruk, A. Gokce, K. Incetan, G. I. Gokceler, M. A. Simsek, K. Ararat, R. J. Chen, N. J. Durr, F. Mahmood, *et al.*, "Endol2h: deep super-resolution for capsule endoscopy," *TMI*, vol. 39, no. 12, pp. 4297–4309, 2020.

[14] W. Chen, Y. Liu, J. Hu, and Y. Yuan, "Dynamic depth-aware network for endoscopy super-resolution," *IEEE JBHI*, vol. 26, no. 10, pp. 5189–5200, 2022.

[15] S. Zou, M. Long, X. Wang, X. Xie, G. Li, and Z. Wang, "A cnn-based blind denoising method for endoscopic images," in *BioCAS*, pp. 1–4, 2019.

[16] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *MedIA*, vol. 68, p. 101900, 2021.

[17] F. Queiroz and T. I. Ren, "Endoscopy image restoration: A study of the kernel estimation from specular highlights," *DSP*, vol. 88, pp. 53–65, 2019.

[18] L. Bai, T. Chen, Q. Tan, W. J. Nah, Y. Li, Z. He, S. Yuan, Z. Chen, J. Wu, M. Islam, *et al.*, "Endouic: Promptable diffusion transformer for unified illumination correction in capsule endoscopy," in *MICCAI*, pp. 296–306, 2024.

[19] F. J. Sánchez, J. Bernal, C. Sánchez-Montes, C. R. de Miguel, and G. Fernández-Esparrach, "Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos," *Machine Vision and Applications*, vol. 28, no. 8, pp. 917–936, 2017.

[20] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE TMI*, vol. 35, no. 2, pp. 630–644, 2015.

[21] S. Mathew, S. Nadeem, and A. Kaufman, "Clts-gan: color-lighting-texture-specular reflection augmentation for colonoscopy," in *MICCAI*, pp. 519–529, 2022.

[22] T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr, "Colonoscopy 3d video dataset with paired depth from 2d-3d registration," *MedIA*, vol. 90, p. 102956, 2023.

[23] G. Yue, D. Cheng, T. Zhou, J. Hou, W. Liu, L. Xu, T. Wang, and J. Cheng, "Perceptual quality assessment of enhanced colonoscopy images: A benchmark dataset and an objective method," *IEEE TCSVT*, vol. 33, no. 10, pp. 5549–5561, 2023.

[24] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *ACM MMSys*, 2017.

[25] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *SData*, vol. 7, no. 1, p. 283, 2020.

[26] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *SData*, vol. 8, no. 1, p. 142, 2021.

[27] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, H. D. Johansen, *et al.*, "Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy," in *MMM*, 2021.

[28] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, *et al.*, "Nerthus: A bowel preparation quality video dataset," in *ACM MMSys*, 2017.

[29] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE TMI*, vol. 35, no. 9, pp. 2051–2063, 2016.

[30] D. Jha, N. K. Tomar, V. Sharma, Q.-H. Trinh, K. Biswas, H. Pan, R. K. Jha, G. Durak, A. Hann, J. Varkey, *et al.*, "Polypdb: A curated multi-center dataset for development of ai algorithms in colonoscopy," *arXiv preprint arXiv:2409.00045*, 2024.

[31] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, *et al.*, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *GIE*, vol. 93, no. 4, pp. 960–967, 2021.

[32] S. Gautam, A. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, and M. A. Riegler, "Kvasir-vqa: A text-image pair gi tract dataset," in *ACM MM-W*, 2024.

[33] S. Gautam, M. A. Riegler, and P. Halvorsen, "Kvasir-vqa-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy," *arXiv preprint arXiv:2506.09958*, 2025.

[34] S. Liu, B. Zheng, W. Chen, Z. Peng, Z. Yin, J. Shao, J. Hu, and Y. Yuan, "Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis," *NeurIPS D&B*, 2025.

[35] B. Khanal, S. Pokhrel, S. Bhandari, R. Rana, N. Shrestha, R. B. Gurung, C. Linte, A. Watson, Y. R. Shrestha, and B. Bhattarai, "Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models," in *MICCAI*, 2025.

[36] F. Mahmood, "A benchmarking crisis in biomedical machine learning," *Nature Medicine*, pp. 1–1, 2025.

[37] A. Koulaouzidis, D. K. Iakovidis, D. E. Yung, E. Rondonotti, U. Kopylov, J. N. Plevris, E. Toth, A. Eliakim, G. W. Johansson, W. Marlicz, *et al.*, "Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes," *EIO*, vol. 5, no. 06, pp. E477–E483, 2017.

[38] T.-H. Hoang, H.-D. Nguyen, V.-A. Nguyen, T.-A. Nguyen, V.-T. Nguyen, and M.-T. Tran, "Enhancing endoscopic image classification with symptom localization and data augmentation," in *ACMMM*, 2019.

[39] S. Ali, N. Ghatwary, B. Braden, D. Lamarque, A. Bailey, S. Realdon, R. Cannizzaro, J. Rittscher, C. Daul, and J. East, "Endoscopy disease detection challenge 2020," *arXiv preprint arXiv:2003.03376*, 2020.

[40] D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, *et al.*, "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *SData*, vol. 10, no. 1, p. 574, 2023.

[41] M. Ye, S. Giannarou, A. Meining, and G.-Z. Yang, "Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations," *MedIA*, vol. 30, pp. 144–157, 2016.

[42] P. Handa, M. Dhir, A. Mahbod, F. Schwarzhans, R. Woitek, N. Goel, and D. Gunjan, "Wcebleedgen: A wireless capsule endoscopy dataset and its benchmarking for automatic bleeding classification, detection, and segmentation," *arXiv preprint arXiv:2408.12466*, 2024.

[43] R. Leenhardt, C. Li, J.-P. Le Mouel, G. Rahmi, J. C. Saurin, F. Cholet, A. Boureille, X. Amiot, M. Delvaux, C. Duburque, *et al.*, "Cad-cap: a 25,000-image database serving the development of artificial intelligence for capsule endoscopy," *EIO*, vol. 8, no. 03, pp. E415–E420, 2020.

[44] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, M. A. Riegler, K. V. Anonsen, *et al.*, "A multi-centre polyp detection and segmentation dataset for generalisability assessment," *SData*, vol. 10, no. 1, p. 75, 2023.

[45] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, pp. 34892–34916, 2023.

[46] C. Niu, Q. Lyu, C. D. Carothers, P. Kaviani, J. Tan, P. Yan, M. K. Kalra, C. T. Whitlow, and G. Wang, "Medical multimodal multitask foundation model for lung cancer screening," *Nature Communications*, vol. 16, no. 1, p. 1523, 2025.

[47] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[48] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *NeurIPS*, vol. 36, pp. 28541–28564, 2023.

[49] E. J. Lai, A. H. Calderwood, G. Doros, O. K. Fix, and B. C. Jacobson, "The boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research," *Gastrointestinal endoscopy*, vol. 69, no. 3, pp. 620–625, 2009.

[50] S. Hattori, M. Iwatate, W. Sano, N. Hasuike, H. Kosaka, T. Ikumoto, M. Kotaka, A. Ichiyanagi, C. Ebisutani, Y. Hisano, *et al.*, "Narrow-band imaging observation of colorectal lesions using nice classification to avoid discarding significant lesions," *WJG*, vol. 6, no. 12, p. 600, 2014.

[51] P. in the Paris Workshop, "The paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to december 1, 2002," *GIE*, vol. 58, no. 6, pp. S3–S43, 2003.

[52] P. J. Pickhardt, K. S. Hain, D. H. Kim, and C. Hassan, "Low rates of cancer or high-grade dysplasia in colorectal polyps collected from computed tomography colonography screening," *Clinical Gastroenterology and Hepatology*, vol. 8, no. 7, pp. 610–615, 2010.

[53] K. W. Schroeder, W. J. Tremaine, and D. M. Ilstrup, "Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis," *New England Journal of Medicine*, vol. 317, no. 26, pp. 1625–1629, 1987.

[54] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang, *et al.*, "Vlmevalkit: An open-source toolkit for evaluating large multi-modality models," in *ACMMM*, 2024.

[55] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Ruhle, L. V. Lakshmanan, and A. H. Awadallah, "Hybrid llm: Cost-efficient and quality-aware query routing," in *ICLR*, 2024.

[56] D. Zhang, J. Wang, and F. Charton, "Only-if: revealing the decisive effect of instruction diversity on generalization," *Preprint*, 2024.

[57] N. Boizard, H. Gisserot-Boukhlef, K. El-Haddad, C. Hudelot, and P. Colombo, "When does reasoning matter? a controlled study of reasoning's contribution to model performance," *arXiv preprint arXiv:2509.22193*, 2025.

[58] S. Lewis-Lim, X. Tan, Z. Zhao, and N. Aletras, "Can confidence estimates decide when chain-of-thought is necessary for llms?," *arXiv preprint arXiv:2510.21007*, 2025.

[59] C. Shi, Y. Su, C. Yang, Y. Yang, and D. Cai, "Specialist or generalist? instruction tuning for specific nlp tasks," in *EMNLP*, 2023.

[60] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *IEEE CVPR*, 2024.

[61] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llavanext: Improved reasoning, ocr, and world knowledge," 2024.

[62] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[63] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[64] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[65] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[66] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.

[67] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," in *NeurIPS*, 2023.

[68] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.

[69] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. H. Chen, X. Wang, R. Zhang, Z. Cai, K. Ji, *et al.*, "Towards injecting medical visual knowledge into multimodal LLMs at scale," in *EMNLP*, 2024.

[70] Y. Liu, Z. Liang, Y. Wang, X. Wu, F. Tang, M. He, J. Li, Z. Liu, H. Yang, S. Lim, *et al.*, "Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly," in *CVPR*, pp. 9087–9097, 2025.

[71] D. Nguyen, M. K. Ho, H. Ta, T. T. Nguyen, Q. Chen, K. Rav, Q. D. Dang, S. Ramchandre, S. L. Phung, Z. Liao, *et al.*, "Localizing before answering: A benchmark for grounded medical visual question answering," in *IJCAI*, 2025.

[72] A. Deng, T. Cao, Z. Chen, and B. Hooi, "Words or vision: Do vision-language models have blind faith in text?," in *IEEE CVPR*, pp. 3867–3876, 2025.

[73] Y. Su, T. Li, J. Liu, C. Ma, J. Ning, C. Tang, S. Ju, J. Ye, P. Chen, M. Hu, *et al.*, "Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning," *arXiv preprint arXiv:2504.01886*, 2025.

[74] Q. Liu, S. Zhang, G. Qin, T. Ossowski, Y. Gu, Y. Jin, S. Kiblawi, S. Preston, M. Wei, P. Vozila, *et al.*, "X-reasoner: Towards generalizable reasoning across modalities and domains," *arXiv preprint arXiv:2505.03981*, 2025.

[75] X. Huang, J. Wu, H. Liu, X. Tang, and Y. Zhou, "Medvlthinker: Simple baselines for multimodal medical reasoning," *arXiv preprint arXiv:2508.02669*, 2025.

[76] Y. Lai, J. Zhong, M. Li, S. Zhao, and X. Yang, "Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models," *arXiv preprint arXiv:2503.13939*, 2025.

[77] Z. Wang, J. Wu, L. Cai, C. H. Low, X. Yang, Q. Li, and Y. Jin, "Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow," in *ICLR*, 2026.

[78] Z. Huang, T. Tang, S. Chen, S. Lin, Z. Jie, L. Ma, G. Wang, and X. Liang, "Making large language models better planners with reasoning-decision alignment," in *ECCV*, pp. 73–90, 2024.

[79] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.

[80] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*, 2024.

[81] Z. Zeng, Z. Zhuo, X. Jia, E. Zhang, J. Wu, J. Zhang, Y. Wang, C. H. Low, J. Jiang, Z. Zheng, *et al.*, "Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence," *arXiv preprint arXiv:2506.02555*, 2025.

[82] W. Zhao, C. Wu, Y. Fan, X. Zhang, P. Qiu, Y. Sun, X. Zhou, Y. Wang, X. Sun, Y. Zhang, *et al.*, "An agentic system for rare disease diagnosis with traceable reasoning," *Nature*, 2026.

[83] Y. Zhang, Y. Huang, Y. Sun, C. Liu, Z. Zhao, Z. Fang, Y. Wang, H. Chen, X. Yang, X. Wei, *et al.*, "Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models," in *NeurIPS D&B*, vol. 37, pp. 49279–49383, 2024.