大模型原理与技术课程设计

姓名：文代霖 学号：BX2524702

一、选题动机与背景

在人工智能技术飞速发展的背景下，大语言模型（LLM）正加速向多模态方向转型，这一趋势正深刻重塑医疗辅助诊断（CAD）的任务范式 。传统的医学影像分析多聚焦于单一模态的特征识别，而现代临床诊断，特别是针对阿尔茨海默症（AD）等复杂疾病，已展现出从单一影像分析向跨模态语义整合转变的紧迫需求 。

临床实践表明，AD 的诊断是一个多维度的综合过程，不仅高度依赖于脑部 3D 核磁共振（MRI）影像所提供的结构化解剖信息，还需深度整合电子健康记录（EHR）中的非结构化或半结构化文本信息，如受试者的年龄、性别、受教育程度、认知功能评估分数（MMSE、CDR、逻辑记忆得分）以及实验室测试结果。这些异构数据相互补充，共同构成了患者健康的完整画像。

然而，现有的视觉语言预训练（VLP）模型（如 CLIP、BLIP 等）及其医学衍生版本（如 MedCLIP、BioViL-T 等）大多聚焦于 2D 图像或 2D 切片，其核心能力往往局限于通用医学问答。面对具有高维度空间特征的 3D 医学影像时，现有模型面临着严峻的挑战：一方面是处理 3D 卷积与海量体素数据带来的计算鸿沟；另一方面是 3D 标注数据相对稀缺导致模型难以充分训练。此外，如何在大规模预训练语言模型中有效嵌入 3D 临床表征，并实现精准的跨模态对齐，仍是该领域亟待解决的核心科学问题。

针对上述瓶颈，本课题选择以 MedAlign-3D 模型为核心，探索一种基于自举（Bootstrapping）策略的新型 CAD 框架。该课题旨在利用创新的 AlignQ-Module 模块，作为连接冻结的 2D 预训练视觉编码器（如 ViT-G）与生物医学大模型（如 BioMedLM）的桥梁。通过引入可学习的线性投影与位置嵌入，该方案能够将 3D 医学影像高效映射至 2D 模型的特征空间，并利用学习到的查询（Queries）实现 3D 影像特征与临床文本语义的深度对齐。这一研究路径不仅显著降低了对大规模标注 3D 数据的依赖及计算成本，更致力于实现具备临床解释性的智能化问答，为构建更高效、更实用的 3D 医疗辅助诊断系统提供理论支撑与工程范式。

二、核心技术与架构实现

MedAlign-3D 的核心创新在于将医学计算机辅助诊断（CAD）任务建模为一个条件文本生成过程，通过构建一个双流编码器与因果语言模型（Decoder）构成的端到端框架，实现 3D 影像、文本描述与诊断问答的深度融合。在该框架下，诊断过程被形式化为：给定受试者的脑部 3D 影像体积I、从电子健康记录（EHR）中提取的自然语言文本描述 T 以及临床问题 Q,模型的目标是生成包含 N 个 Token 的诊断答案A。其优化的数学本质是最大化条件对数似然概率，即通过调整模型参数θ使得下式最优：

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log P(A_i | A_{<i}, I, T, Q; \theta)$$

在视觉特征处理层面，为了克服3D医学影像与预训练2D视觉模型之间的维度不匹配，系统引入了 AlignQ-Module。该模块首先将输入的 3D 影像体积I划分为一系列 3D 子体积$\{I_{v_i}\}$，并通过线性投影函数f_{φ_1}将其映射为 1D 影像嵌入。为了保留解剖结构信息，系统加入了可学习的位置嵌入f_{φ_2}，随后将其输入冻结的预训练 2D 视觉编码器（如 ViT-G/14）中，提取出多尺度的视觉特征$f_i = \left\{ f_\phi \left(f_{\varphi_1}(I_{v_i}), f_{\varphi_2}(I_{v_i}) \right) \right\}$。

随后，AlignQ-Module 利用一组L个可学习的查询作为媒介，通过自注意力机制实现查询间的特征交互，并利用交

叉注意力机制从视觉编码器的输出中提取关键解剖特征。这些查询最终被映射到语言模型的潜在空间中，形成视觉前缀H_v。在语言编码流中，文本描述T、问题Q和答案A经过标准分词处理后转换为嵌入向量。为了引导语言模型进行有效的推理，系统构建了一个结构化的 Prompt 模板，其排列顺序通常为视觉特征H_v、文本描述H_T、问题 Token H_Q，最终由语言模型生成答案序列A。

在模型训练与参数更新策略上，为了在有限的医学数据下保持通用语言模型的推理能力并防止过拟合，研究采用了参数高效微调（PEFT）技术。除了保持语言模型核心权重冻结外，还在注意力机制的查询矩阵Q_w和数值矩阵V_w中嵌入了低秩适配器。整个系统的损失函数由两部分组成：一是特征对齐损失$\mathcal{L}_{FA}$，它基于图像-文本对比学习（ITC）机制，强化视觉查询与诊断文本之间的语义关联；二是语言生成损失$\mathcal{L}_{LG}$，采用交叉熵评估生成答案与标准答案的一致性。最终的总损失函数定义为：

$$\mathcal{L}_{total} = \mathcal{L}_{FA} + \lambda_{LG}\mathcal{L}_{LG},$$

其中λ_{LG}为平衡两个任务权重的超参数。这种方法不仅实现了模态间的高效对齐，还赋予了模型在零样本场景下通过思维链（CoT）进行医学推理的能力。

三、实验评估与学术贡献

对 ADNI、NACC、OASIS、AIBL、MIRIAD 五个数据集的大部分图像进行预训练后，在两个任务上评估了 MedAlign-3D：（1）零样本分类，即直接应用预训练的 MedAlign-3D 对来自 AIBL 和 MIRIAD 数据集的未见受试者进行三类分类，即正常对照（NC）、轻度认知障碍（MCI）和阿尔茨海默病（AD）；（2）零样本医学视觉问答，根据输入的图像和文本描述，为未见过的 AIBL 或 MIRIAD 受试者生成初步诊断，并提供做出该决策的一些理由。

表 1. AD 数据集的人口统计统计。F：女性，M：男性，Educ：教育水平，SES：社会经济地位，MMSE：简易精神状态检查，CDR：临床痴呆评分，LM：逻辑记忆，E/L/S/PMCI：早期、晚期、稳定及渐进性轻度认知障碍，IMCI：受损但非轻度认知障碍，DEM：痴呆。

Datasets		ADNI	NACC	OASIS	AIBL	MIRIAD
Images		10287	15354	3020	1002	708
Text	F/M	4710/5677	9058/6296	1798/1222	471/531	393/315
	Age	45-95	19-102	18-98	42-96	55-87
	Educ	9860	15329	2300	-	-
	SES	-	-	2153	-	
	MMSE	9385	7867	2293	1002	268
	CDR	9401	15354	2300	1002	46
	LM	7189	7654	-	1002	-
	Diagnosis	NC, MCI, AD, E/L/S/PMCI	NC, IMCI, MCI, DEM	DEM, Non-DEM	NC, MCI, AD	NC, AD

1、零样本分类：

表 2 展示了 MedAlign-3D 与多种基线方法在五个数据集上的定量对比结果，分类性能通过五次实验运行的平均准确率（ACC%）进行衡量。该实验深入对比了三种主流的大型语言模型，包括参数量为 3.4B 的 FLAN-T5、1.5B 的 BioGPT 以及 2.7B 的 BioMedLM，并分别针对“仅文本”、“冻结语言模型”以及“LoRA 微调”三种不同设置进行了严谨的性能评估。结果表明，即使在完全冻结语言模型参数的预训练模式下，引入 3D 视觉模态后的模型表现较仅使用文本的模型仍实现了 14.0%至 44.8%的显著提升，这充分证明了 3D 影像扫描在阿尔茨海默病诊断中的必要性。在三个基准语言模型中，BioMedLM 在捕捉提示词与医学知识间的依赖关系方面表现最为出色。通过结合 LoRA 参数高效微调技术，系统性能在冻结设置的基础上进一步实现了 1.3%到 13.5%的增长，最终使得基于 BioMedLM 架构并应用 LoRA 微调的 MedAlign-3D 系统在包括零样本测试集 AIBL（准确率达 80.8%）和 MIRIAD（准确率达 71.0%）在内的所有数据集上均取得了最优的诊断准确率。

表 2. MedAlign-3D 与基线方法在五个数据集上的定量比较。分类性能通过五次运行的平均准确率 (ACC(%))来衡量。最佳分数以粗体显示。(†：零样本)

Methods		LM size	Learnable params	ADNI -3x200	NACC -3x200	OASIS -2x200	AIBL†	MIRIAD†
FLAN-T5	Text only	3.4B	-	37.0%	39.5%	46.7%	33.3%	60.0%
Ours w/T5	Frozen		151M	50.5%	69.2%	61.3%	54.7%	64.0%
	LoRA		156M	64.0%	77.3%	75.8%	59.2%	66.8%
BioGPT	Text only	1.5B	-	25.7%	21.7%	28.3%	26.7%	50.0%
Ours w/BioGPT	Frozen		151M	56.3%	66.5%	66.0%	60.7%	55.2%
	LoRA		156M	62.2%	72.3%	71.7%	62.4%	59.7%
BioMedLM	Text only	2.7B	-	62.5%	63.5%	61.8%	65.7%	46.3%
Ours w/BioMedLM	Frozen		151M	71.2%	82.0%	79.8%	77.8%	66.1%
	LoRA		154M	**78.7%**	**83.3%**	**85.3%**	**80.8%**	**71.0%**

图 1（a-c）可视化了从 AIBL 数据集中抽取的未见过的受试者的零样本分类过程。以图 1（b）为例，尽管该受试者的文字描述与健康受试者没有显著差异，例如 CDR 分数为 0；但在脑部扫描中，海马和脑室显示存在异常萎缩。MedAlign-3D 为该受试者提供了正确的轻度认知障碍（MCI）诊断。

2、零样本医学视觉问答：

图 1（d-f）定性地展示了 MedAlign-3D 在 AIBL 数据集上的零样本医学 VQA 能力。在一个简单的基于分类的问题之后，MedAlign-3D 提供了一种自然的方式来执行 VQA 并展示思路链。虽然由于各种原因，包括大语言模型中的医学知识不准确、激活了错误的推理路径，或缺乏关于新图像内容的最新信息，MedAlign-3D 可能会生成不理想的答案。

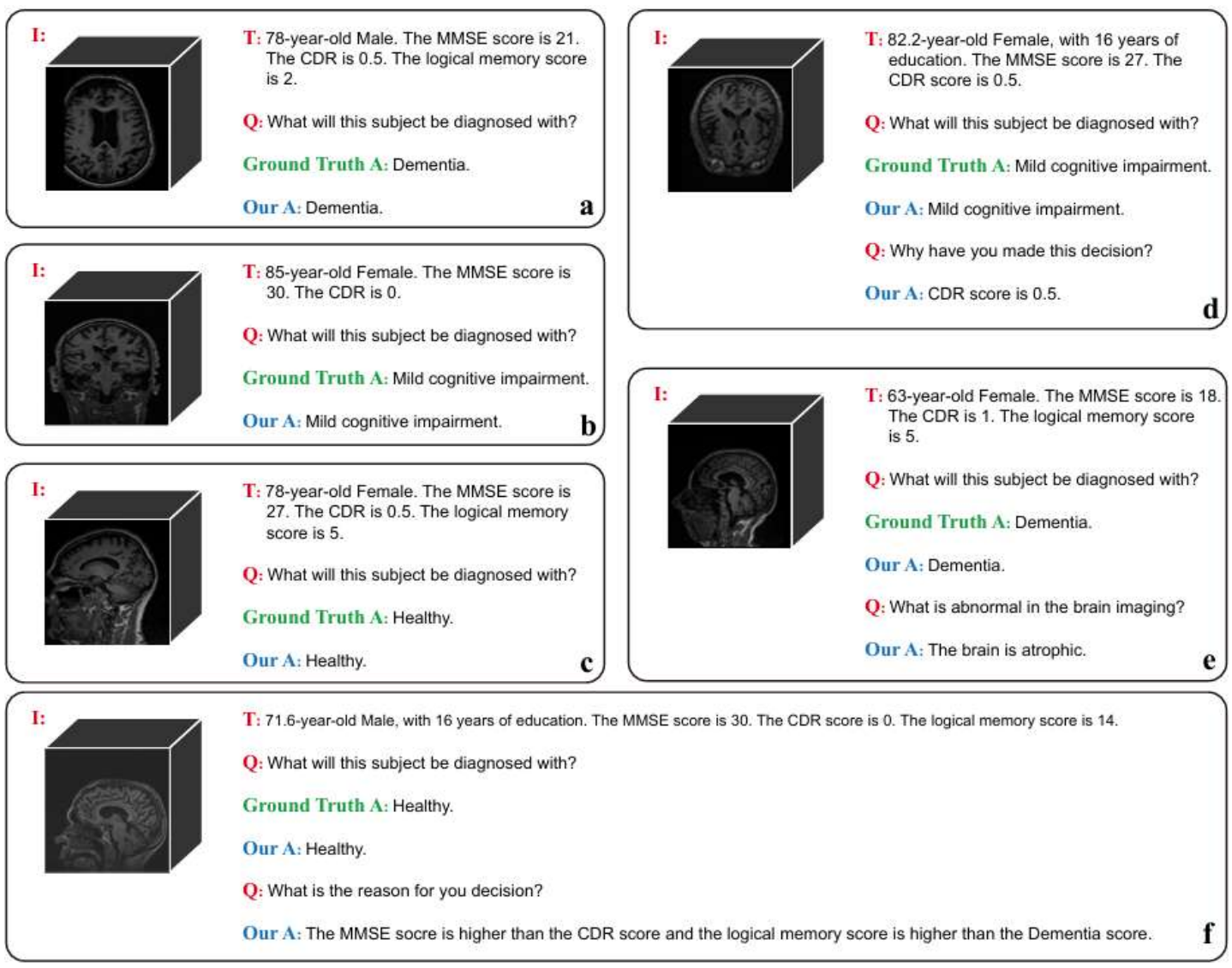

图 1. 在 AIBL 数据集上的零样本结果示例，这些结果是由基于 BioMedLM 并经过 LoRA 微调的 MedAlign-3D 生成的。

表 3 展示了 MedAlign-3D 在 M3D-VQA-AD 闭口问答数据集上的零样本医学视觉问答性能评估结果。该评估旨在测试模型在面对与阿尔茨海默病（AD）相关的临床问题及给定选项时，直接从 3D 影像中提取关键特征并选择正确答案的能力。实验结果显示，MedAlign-3D 取得了 77.96%的准确率，相较于基准模型 M3D-LaMed 的 72.88%准确率，实现了显著的性能超越。这一数据有力地证明了 MedAlign-3D 在零样本场景下，能够精准地将 3D 医学影像的视觉表征与复杂的临床问答语义进行深度对齐。

表 3. MedAlign-3D 在 M3D-VQA-AD 封闭式数据集上的零样本医学视觉问答结果。

Methods	Accuracy
M3D-LaMed	72.88%
Ours	**77.96%**

图 4 展示了 MedAlign-3D 在 M3D-VQA-AD 数据集上进行零样本医学视觉问答的定性结果示例。在该任务中，问题和选项被作为提示输入，模型需要根据 3D 影像信息从给定选项中选择正确答案。在图 4 展示的第一个案例中，问题询问“支持阿尔茨海默病诊断的特征是什么？”。选项包括：A.脑室周围及深部白质低密度、B.脑室扩大、C.颅骨骨折、D.颅内出血。MedAlign-3D 生成的答案准确选中了选项 A，与标准答案完全一致。在第二个案例中，针对“阿尔茨海默病典型的发病年龄范围是多少？”这一问题，模型从四个年龄段选项（20s-30s 到 80s-90s）中正确选择了“C.60s-70s”。这一预测结果同样与标准答案相符。这些示例直观地证明了 MedAlign-3D 能够有效地将 3D 医学影像特征与文本中的医学知识及特定选项进行对齐，从而在未见过的任务中实现准确的推理和判断。

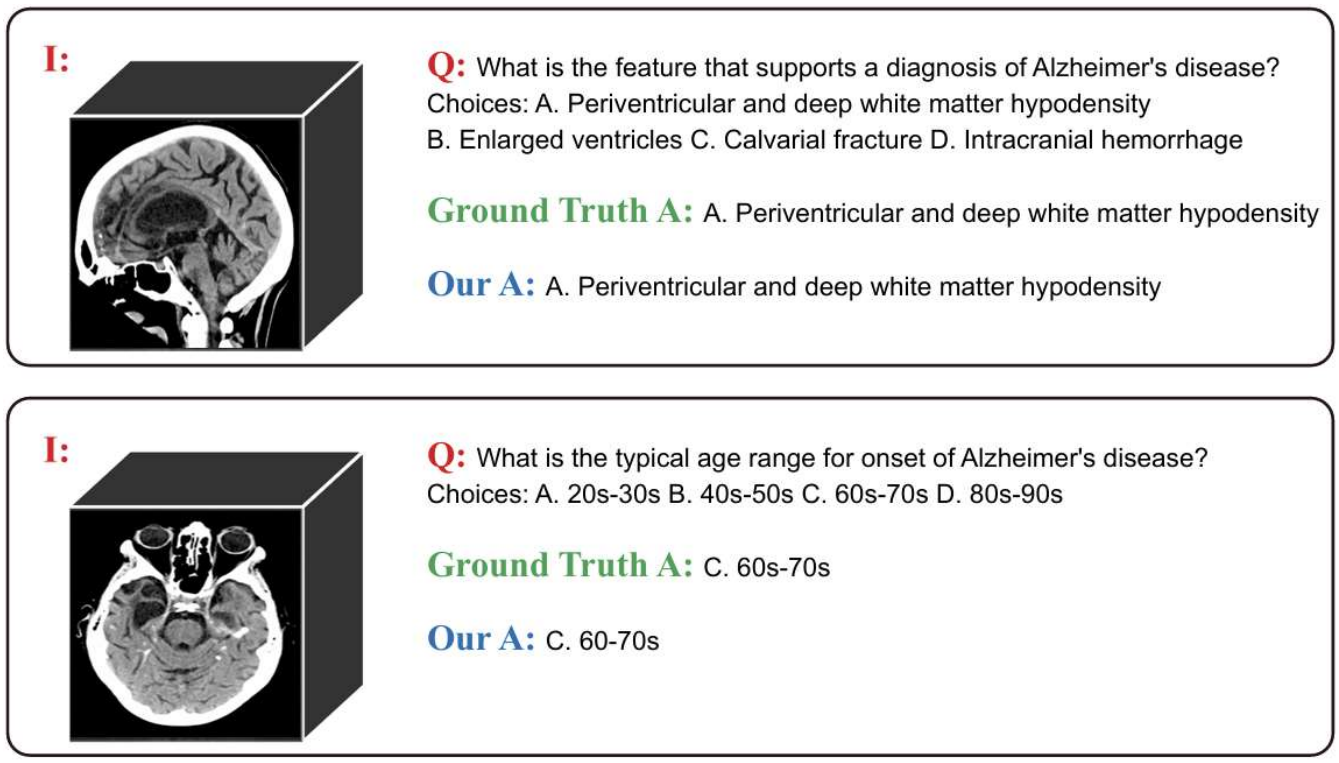

图 2.在 M3D-VQA-AD 数据集上的零样本医疗 VQA 结果示例。

3、学术贡献：

本课设提出了一种名为 MedAlign-3D 的轻量级多模态计算机辅助诊断系统，旨在通过整合 3D 医学影像与电子健康记录中的文本信息，攻克阿尔茨海默病的高精度辅助诊断难题。该系统通过创新的 AlignQ-Module 模块，成功桥接了预训练的冻结 2D 视觉编码器与大型语言模型，有效解决了 3D 医疗影像在维度对齐与特征提取上的挑战。依托于包含 30,000 余份影像体积的大规模数据集，MedAlign-3D 在零样本分类及医疗视觉问答任务中展现了卓越的性能，能够针对未见过的受试者生成准确的初步诊断并提供具有临床逻辑的解释理由。