MIRAGE: Multimodal Integration via Representational Alignment using Graphormer and Encoders

引言（Introduction）

心血管疾病（Cardiovascular Diseases, CVDs）仍然是全球范围内的主要死亡原因之一，约占全球死亡总数的三分之一。其中，心肌梗死（Myocardial Infarction, MI，俗称心脏病发作）构成了严重的公共健康挑战。MI通常由于冠状动脉血流受阻而引发，这种阻塞往往是动脉堵塞所致，因而需要迅速且准确的诊断以实现有效干预。及时的检测不仅对于降低死亡率至关重要，同时也有助于减轻如心力衰竭等严重并发症，从而改善患者的整体预后。

心电图（Electrocardiogram, ECG）因其实时监测能力、可及性和稳定的诊断价值而被广泛认可为一种基本的、非侵入性且成本低廉的诊断工具[1]。尽管ECG被广泛应用，传统的ECG分析方法——尤其是依赖单一模态的方法——仍存在显著局限。这类方法易受噪声干扰、难以适应不同患者间的个体差异，并常常缺乏对系统性生理状态的整体理解。近年来，深度学习技术的快速发展极大提升了ECG分类的性能，但现有的大多数模型通常仅依赖于时间序列数据或静态图像中的一种。这种单一模态的建模方式忽视了多模态融合所带来的协同效应，从而在诊断准确性与可解释性方面受到限制。

在此背景下，多模态学习逐渐成为医学诊断中的研究热点。例如，实验室检测提供了ECG无法获取的关键生化背景信息；而ECG图像则体现了结构与形态学特征，为视觉诊断提供重要线索。这些互补模态的集成能够形成更丰富、更具信息性的患者表征，从而支持更精确、个性化的临床评估。例如，Toprak 等人[4]展示了基于高敏肌钙蛋白实验数据训练的深度学习模型在MI分诊中的优越性能，验证了生物标志物在心血管急症诊断中的价值。此外，Kilimci 等人[5]指出，基于Transformer的ECG图像模型可有效识别心脏疾病，进一步证明了视觉线索在自动化诊断中的潜力。

值得注意的是，多模态方法不仅在心血管领域中取得了成果。在其他研究中，Tang 等人[6]提出了一种用于少样本临床问答的元学习ECG-语言模型，强调了跨模态融合对于泛化能力的提升；Holmstrom 等人[7]则指出，ECG信号甚至可以揭示非心脏疾病如慢性肾病等共病情况，进一步体现了ECG潜在的信息价值。这些研究激发了跨越传统波形分析范畴的多模态方法的研究热情。

受到这些临床与方法学进展的启发，本文提出了一种新颖的多模态框架——**MIRAGE**（Multimodal Integration via Representational Alignment using Graphormer and Encoders），该模型专为基于ECG的心血管疾病分类任务设计。MIRAGE协同融合了三类互补的数据模态：反映电活动的12导联ECG时间序列、展现形态学模式的ECG灰度图像，以及体现系统性生化状态的实验室检测数据。

MIRAGE中的架构设计有针对性地解决了关键的临床挑战。例如，传统的时间序列模型难以捕捉12导联ECG数据中潜在的导联间依赖关系，为此本文引入了Graphormer编码器，将每个导联视作图节点，从而显式建模其生理关联性，提升诊断的准确性与可解释性。

与此同时，针对ECG图像的分析部分采用了视觉Transformer（Vision Transformer, ViT），以反映临床实践中对波形形态的视觉评估方法。与传统卷积方法不同，ViT具备全局自注意力机制，能更有效地捕捉如ST段抬高等形态特征[5]。

为进一步增强诊断能力，MIRAGE在两个编码分支中均引入了特征层级线性调制（Feature-wise Linear Modulation, FiLM）机制。该机制利用肌钙蛋白、D-二聚体等实验室生物标志物作为调制信号，动态调整特征表示以反映个体的生理状态[4]，实现具有上下文感知能力的个性化表征提取。

为了综合整合上述多模态信息，MIRAGE设计了一个基于Transformer的融合模块，能够弥合不同模态间的语义差异，建模时间序列、图像与生化信息间复杂的交互。此外，模型还引入了一种对比学习策略，与标准分类任务并行训练，以鼓励跨模态表示的一致性与对齐。

综上，MIRAGE 的主要贡献包括：

- 构建了结合Graphormer与ViT的双分支编码器，有效挖掘ECG的时间与空间信息；
- 引入基于FiLM的生物标志物调制机制，实现个性化、具生理背景的特征提取；
- 构建了一个Transformer驱动的融合策略，实现多模态信息的深度整合；
- 在训练过程中结合了对比学习机制，增强了多模态表示的一致性与鲁棒性。

通过上述创新设计，MIRAGE有效克服了传统单模态ECG方法的诸多限制，为心肌梗死的自动化诊断提供了一种鲁棒、可解释且具临床实用价值的解决方案。

相关研究（Related Work）

A. 心电图分类：进展与挑战

自动化心电图（ECG）分类因其在心血管疾病诊断中的关键作用而受到广泛关注。传统的机器学习方法主要依赖人工特征提取与浅层分类器。随着深度学习的发展，卷积神经网络（CNN）和循环神经网络（RNN）成为主流方法，在MIT-BIH和PTB-XL等标准数据集上取得了显著成绩。近年来，Transformer架构因其对全局时间依赖性的建模能力优越，在ECG任务中也获得越来越多应用【8】。

然而，大多数现有模型依然是**单模态**，仅限于处理原始ECG信号或图像数据。这种单一模态的策略常常忽视了临床诊断的多维信息结构，例如多种数据源与生物标志物的综合判断。例如，Liu 等人【9】提出了基于CNN的急性心肌梗死检测模型，但未引入生化背景；Jafarian 等人【10】研究了浅层与深度模型用于MI定位，但仅依赖于信号形态特征；Lee 等人【11】使用深度学习从ECG中筛查左心室射血分数低的患者，却未探索模态融合。在更广泛的背景下，Kalmady 等人【12】开发了可预测15种心血管疾病的深度学习模型，基于超过160万条ECG记录，展示了ECG在多病种大规模筛查中的潜力【13】。

此外，近期研究也突出了ECG在检测系统性疾病方面的潜力。例如，Holmstrom 等人【7】报道其模型在检测慢性肾病方面表现优异，特别是在年轻人群中效果显著；Khurshid 等人【14】则发现将ECG深度学习与临床特征结合有助于提升房颤风险预测准确性。

B. 图神经网络与Transformer在医学中的应用

图神经网络（GNN）在医学结构化数据建模中展现出巨大潜力。就ECG而言，相较于传统序列模型，GNN能更显式地捕捉导联之间的交互关系。例如，Backhaus 等人【15】利用人工智能对心肌应变进行量化，证明了心脏特征中的空间依赖性具有临床价值。

与此同时，视觉Transformer（ViT）与时间序列Transformer因其建模全局模式的能力而在医疗领域获得关注。Nie 等人【8】将Transformer应用于长时间ECG序列预测中，展示了其在时间建模中的优势；Kilimci 等人【5】则表明ViT变体在ECG图像分类任务中优于传统CNN。但目前，较少研究将GNN与Transformer统一于同一框架之中。MIRAGE正是通过将Graphormer与ViT整合，联合学习结构与形态特征，填补了这一空白。

C. 心血管诊断中的多模态融合

多模态整合通过融合不同模态的互补信息，有望显著提升诊断的可靠性。已有若干研究探索了将ECG与临床参数结合的路径。例如，Lampert 等人【16】使用深度学习预测室性早搏患者的心肌病，但所用的结构化临床信息较为有限；Yuan 等人【17】证明ECG在预测房颤中的价值，但仅使用窦性心律数据；Wei 等人【18】提出了BMIRC模型，通过掩码自编码器融合ECG时域与频域数据，增强了跨模态表征学习能力。然而，该模型仅限于两种信号模态，并未利用外部临床信息。

Toprak 等人【4】进一步验证了高敏肌钙蛋白（hs-cTn）数据在深度学习心肌梗死预测中的临床意义；Boeddinghaus【19】与 Doudesis 等人【20】也报道了机器学习驱动的ECG解读在急性MI分诊中的优越性，相较于传统规则路径更为准确。Al-Zaiti 等人【21, 22】强调了将ECG特征与临床元数据结合的重要性，可用于风险分层与闭塞性MI识别。类似地，Sun 等人【13】基于160万条加拿大ECG记录，验证了引入实验室信息后模型对群体死亡率预测的能力，说明深度模型可服务于大规模实时风险分层。

MIRAGE在此基础上进一步推进，首次整合ECG时间序列、ECG图像与实验室特征，并引入基于FiLM的病人特征调制机制，从而实现更具个性化与可解释性的表示学习，契合真实临床工作流程。

D. 表示对齐的对比学习方法

对比学习方法在自监督或有监督框架下展现出强大的多模态特征对齐能力。在ECG研究中，对比学习仍属新兴方向。Kowlgi 等人【23】虽将深度学习用于PVC-心肌病检测，但并未涉及跨模态对比对齐；Wei 等人【18】通过引入对比目标提高了跨域泛化能力，为多模态应用奠定了基础。

近期如TimeMAE与TimesURL等模型进一步扩展了对比学习的边界。TimeMAE采用窗口化预测任务的解耦掩码自编码器，提升了时间序列表示的保真度【24】；TimesURL则结合频率-时间增强策略与双重Universum负样本，联合优化对比与重构目标，适用于多种下游任务【25】。

MIRAGE进一步发展这一方向，设计了双向对比损失函数，在ECG时间序列与图像分支之间实现表示对齐，促进了模态间的一致融合，同时保留各自的语义特征。

小结

MIRAGE的提出受到了以下几方面研究成果的启发并加以拓展：

- ECG分类方法【9–11, 17】；
- 结构建模技术【8, 15】；
- 多模态融合策略【4, 12, 16–22】；
- 对比学习机制【18, 23】。

MIRAGE最终构建了一个统一的、具病人意识的多模态学习框架，用于实现鲁棒的心血管疾病预测。

方法（Method）

我们提出了 MIRAGE（Multimodal Integration via Representational Alignment using Graphormer and Encoders），一个统一的多模态学习框架，专为基于 ECG 的疾病预测任务设计。MIRAGE 集成了三种异构模态：12 导联的 ECG 时间序列、ECG 灰度图像，以及实验室检验特征。整体架构包含三大核心组件：

1. 一个带有 FiLM 增强机制的 Graphormer 模块，用于从 ECG 信号中建模时间与导联间的结构模式，同时注入临床上下文；
2. 一个同样通过 FiLM 调制的 Vision Transformer（ViT）图像编码分支，用于提取 ECG 图像中的形态特征；
3. 一个基于 Transformer 编码器的跨模态融合模块，用于联合优化来自多模态的表示。

此外，MIRAGE 引入了对比学习策略，在信号与图像模态之间显式对齐表示，从而获得更稳健、具上下文感知的融合表示，用于最终分类。

本框架的设计旨在解决现有单模态或弱融合模型的不足，实现时间域、空间域及临床特征的深度交互。

A. 数据与特征提取

传统的 ECG 分类模型大多依赖单一模态，如时间序列信号或波形图像，并忽略了来自实验室检测等补充信息的临床洞见。MIRAGE 通过融合以下三种互补模态构建更全面的患者健康表征：

$$\mathcal{D}_i = \{\mathbf{X}_{\text{ECG},i}, \mathbf{X}_{\text{img},i}, \mathbf{x}_{\text{lab},i}, y_i\}$$

其中：

- $\mathbf{X}_{\text{ECG},i} \in \mathbb{R}^{12\times T}$：12 导联 ECG 时间序列；
- $\mathbf{X}_{\text{img},i} \in \mathbb{R}^{1\times 224\times 224}$：ECG 灰度图像；
- $\mathbf{x}_{\text{lab},i} \in \mathbb{R}^{D_{\text{lab}}}$：实验室特征向量；
- $y_i \in \{0, 1\}$：二分类标签。

每种模态在心血管医学中具有明确的互补性与生理可解释性：

- **ECG 时间序列**：提供高分辨率电生理活动，12 导联分别从不同解剖角度记录，能检测诸如 ST 段抬高、T 波倒置等异常。
- **ECG 图像**：保留波形形态、幅度关系及节律模式，符合临床视觉解读习惯，为模型提供形态结构信息。
- **实验室特征**：反映系统性生理状态，如肌钙蛋白（心肌损伤）、D-二聚体（血栓）、白细胞计数（炎症）等，有助于在 ECG 表现相似时做出区分。

通过集成三模态，MIRAGE 提升了表示的鲁棒性与语义丰富度，同时减少了噪声干扰、数据缺失及模糊波形所带来的误判风险。

MIRAGE 的一项核心创新在于**显式引入实验室数据**作为 FiLM 的调制输入，这是多数公开 ECG 数据集所缺乏的。该机制支持对 ECG 表示进行个性化调整。例如，两名患者 ECG 形态相似但肌钙蛋白水平差异显著，MIRAGE 可据此差异调节表示，从而提高识别准确率。

为实现这种多模态融合，我们构建了一个具备三模态同步采集的数据集，使时间、空间与生化信息在样本级精准对齐，提升模型整体一致性与临床相关性。

B. 时间序列分支：Graphormer 与 FiLM

为捕捉 12 导联 ECG 信号中的时空结构特性，MIRAGE 采用基于 Graphormer 的编码器，并结合 Feature-wise Linear Modulation（FiLM）机制以引入病人特异性上下文。

设 $\mathbf{X}_{\text{ECG}} = [\mathbf{x}_1, \ldots, \mathbf{x}_{12}] \in \mathbb{R}^{12\times T}$，其中每个 $\mathbf{x}_\ell \in \mathbb{R}^T$ 表示单个导联的时间序列。首先通过线性变换投影至潜在空间：

$$\mathbf{h}_\ell^{(0)} = \mathbf{W}_{\text{proj}} \cdot \mathbf{x}_\ell + \mathbf{b}_{\text{proj}}, \quad \forall \ell = 1, \ldots, 12$$

接着，引入 Graphormer 模块进行导联间结构建模，其核心为引入可学习的空间偏置 $B \in \mathbb{R}^{12\times 12}$，用于自注意力机制：

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top + \mathbf{B}}{\sqrt{d}}\right) \cdot \mathbf{V}$$

为实现个性化调制，MIRAGE 引入 FiLM 层，根据实验室特征生成缩放与偏移参数：

$$\gamma = f_\gamma(\mathbf{x}_{\text{lab}}), \quad \beta = f_\beta(\mathbf{x}_{\text{lab}})$$

用于调制中间层输出：

$$\mathbf{H}_{\text{FiLM}}^{(l)} = \gamma \odot \mathbf{H}^{(l)} + \beta$$

最终，我们在序列前加入一个可学习的 [CLS] 向量，经过 Transformer 编码后提取全局表示：

$$\mathbf{z}_{\text{ts,cls}} = \text{TransformerEncoder}([\mathbf{h}_{\text{cls}}; \mathbf{H}^{(L)}])$$

C. 图像分支：ViT 与 FiLM

图像分支基于 Vision Transformer（ViT）架构，并同样引入 FiLM 调制机制，以在视觉表征中引入实验室上下文。

给定 ECG 灰度图像 $\mathbf{X}_{\text{img}} \in \mathbb{R}^{1\times H\times W}$，将其划分为 N 个大小为 $P \times P$ 的 patch，并进行展平与线性变换：

$$\mathbf{z}_i^{(0)} = \mathbf{W}_{\text{patch}} \cdot \text{Flatten}(\mathbf{X}_i) + \mathbf{b}_{\text{patch}}, \quad i = 1, \ldots, N$$

添加可学习的分类 token 与位置编码：

$$\mathbf{Z}^{(0)} = [\mathbf{z}_{\text{cls}}; \mathbf{z}_1^{(0)} + \mathbf{e}_1; \ldots; \mathbf{z}_N^{(0)} + \mathbf{e}_N]$$

FiLM 调制依旧基于实验室特征生成 $\gamma^{(l)}, \beta^{(l)}$，作用于每一层输出：

$$\mathbf{Z}_{\text{FiLM}}^{(l)} = \gamma^{(l)} \odot \mathbf{Z}^{(l)} + \beta^{(l)}$$

最终从分类 token 位置提取图像分支表示：

$$\mathbf{z}_{\text{img,cls}} = \mathbf{Z}_0^{(L)}$$

D. 融合模块：Transformer Encoder

为整合时间序列与图像表示，MIRAGE 引入基于 Transformer 的融合模块，建模模态内与模态间的全局语义依赖。

融合输入包括：

- $\mathbf{z}_{\text{ts,cls}}, \mathbf{z}_{\text{img,cls}}$：两条主分支的全局 token；
- $\mathbf{H}_{\text{ts}}, \mathbf{H}_{\text{img}}$：导联级与图像 patch 级表示。

构建融合 token：

$$\mathbf{z}_{\text{global}} = \mathbf{W}_f[\mathbf{z}_{\text{ts,cls}}; \mathbf{z}_{\text{img,cls}}] + \mathbf{b}_f$$

拼接所有序列形成输入：

$$\mathbf{Z}_{\text{fusion}}^{(0)} = [\mathbf{z}_{\text{global}}; \mathbf{H}_{\text{ts}}; \mathbf{H}_{\text{img}}]$$

送入多层 Transformer 进行融合建模：

$$\mathbf{Z}_{\text{fusion}}^{(L)} = \text{TransformerEncoder}(\mathbf{Z}_{\text{fusion}}^{(0)})$$

提取融合后的主表示：

$$\mathbf{z}_{\text{fusion,cls}} = \mathbf{Z}_{\text{fusion}}^{(L)}[0]$$

E. 对比学习：模态对齐

为实现模态一致性，MIRAGE 引入对比学习，拉近同一病人在不同模态下的表示，同时区分不同病人样本。

对比损失基于对称 InfoNCE 构建，设 batch 中第 i 个样本的两个模态表示为 $\mathbf{z}_{\text{ts},i}$ 和 $\mathbf{z}_{\text{img},i}$，则损失为：

$$\mathcal{L}_{\text{con}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\ell_i^{\text{ts}\to\text{img}} + \ell_i^{\text{img}\to\text{ts}} \right)$$

其中：

$$\ell_i^{\text{ts}\to\text{img}} = \log \frac{\exp(\text{sim}(\mathbf{z}_{\text{ts},i}, \mathbf{z}_{\text{img},i})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{z}_{\text{ts},i}, \mathbf{z}_{\text{img},j})/\tau)} \quad \ell_i^{\text{img}\to\text{ts}} = \log \frac{\exp(\text{sim}(\mathbf{z}_{\text{img},i}, \mathbf{z}_{\text{ts},i})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{z}_{\text{img},i}, \mathbf{z}_{\text{ts},j})/\tau)}$$

其中 $\text{sim}(\cdot,\cdot)$ 为余弦相似度，τ 为温度参数。

F. 分类与训练目标

最终的分类由融合表示 $\mathbf{z}_{\text{fusion,cls}}$ 进行：

$$\hat{y} = \text{softmax}(\mathbf{W}_{\text{cls}} \cdot \mathbf{z}_{\text{fusion,cls}} + \mathbf{b}_{\text{cls}})$$

分类损失 $\mathcal{L}_{\text{cls}}$ 使用交叉熵计算。

总损失函数整合分类与对比损失：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{con}}$$

其中 $\lambda_{\text{cls}}, \lambda_{\text{con}}$ 控制两者权重。

G. 方法小结

MIRAGE 构建了一个融合时间序列 ECG、图像 ECG 与实验室特征的统一多模态框架。通过结合模态特定编码器、FiLM 调制与 Transformer 融合模块，以及对比学习机制，模型能够捕捉复杂的时空-生理交互特征，提升心肌梗死诊断的准确性、鲁棒性与临床可解释性。