研究报告

基础 ID: pdf-2506.10006v2*20260416025902 **来源**: ACM MM'25 论文 — 双模态灵活的 HER2 表达预测框架 **文献结果**:* `*data/lit*results/pdf-2506.10006v2_20260416025902/lit.md` **报告日期**: 2026-04-16

1. 执行概述

HER2（人表皮生长因子受体 2）在乳腺癌组织病理学中的表达预测是一项高风险的临床任务，对治疗决策具有重要影响。正在审查的 ACM MM'25 论文介绍了一个深度学习框架，该框架在双模态组织病理学输入（H&E 和 IHC 染色图像）下实现了 95.09% 的准确率，并在仅有 H&E 图像可用时保持 94.25% 的准确率，利用交叉模态生成对抗网络（CM-GAN）从 H&E 重建缺失的 IHC 模态。这种双模态灵活性解决了一个关键的实际限制：在许多临床环境中，尤其是资源有限的环境中，通常由于成本、组织可用性或实验室工作流程的限制，无法获得同一患者组织块的配对 H&E 和 IHC 图像。

文献分析表明，这项工作处于计算病理学三个活跃研究方向的交汇点：多模态深度学习中的缺失模态鲁棒性、组织病理学的跨模态图像合成以及可解释的 HER2 评分自动化。基于实证分析的最重要发现是，该框架的跨模态重建质量（SSIM = 0.39–0.51）在绝对值上相对较低——然而下游分类准确率仍然很高——这表明分类头对合成伪影的鲁棒性超出了人类病理学家的能力（BCI 病理学家在解释生成的 IHC 图像时仅获得 37.5–40% 的准确率，见 P012）。第二个最重要的发现是，病理信息解耦编码器的共享特征分解方法得到了 MODES（P008，npj Digital Medicine 2025）的独立验证，这验证了在完全不同的临床领域（心血管成像与 ECG 和心脏 MRI）中解耦原则的有效性，显著增强了架构设计的合理性。

最关键的未解决问题是单一机构评估。66 个实验室的多中心染色标准化基准研究（P010，Scientific Reports 2026）提供了具体证据，表明即使是同一组织块在不同实验室染色也会产生显著的视觉差异，这使得跨机构的泛化成为部署中的一个非平凡问题。此外，该框架未提供不确定性量化，这对于临床决策支持系统几乎是必需的。HER2-low（IHC 1+）边界——最具治疗重要性和诊断上最困难的类别——仅通过总体准确率隐含评估，未报告每类的细分。

2. 问题设定和来源背景

2.1 临床背景：乳腺癌中的 HER2 表达

HER2 是乳腺癌细胞上的一种蛋白质受体，当其过表达（HER2 阳性，IHC 3+ 或 IHC 2+ 且 ISH 阳性）时，表明肿瘤行为具有侵袭性，并符合使用靶向治疗（如曲妥珠单抗（Herceptin）和抗体药物偶联物 T-DXd）的资格。最近，HER2-low 表达（IHC 1+ 或 IHC 2+ 且 ISH 阴性）获得了治疗意义，因为这些肿瘤对 T-DXd 有反应，尽管它们以前被认为是 HER2 阴性。这一临床发展使得准确的 HER2 评分——特别是 HER2-low 边界——比

以往任何时候都更加重要。评分标准从 HER2 0（无表达）、HER2 1+（弱表达）、HER2 2+（中等表达，需要 ISH 确认）到 HER2 3+（强过表达）。

标准的诊断工作流程包括 H&E（苏木精-伊红）染色以进行一般形态评估，随后进行 IHC（免疫组化）染色，专门突出细胞膜上的 HER2 蛋白表达。在常规临床实践中，H&E 和 IHC 图像很少配对用于同一组织块，其可用性取决于不同的组织切片、染色运行和实验室日程。

2.2 计算病理学中的缺失模态问题

现有的计算病理学多模态 AI 模型通常假设在推理时可用完整的模态——即来自多种成像模态的配对输入可用。这一假设在实际临床工作流程中失败的原因有三个：（1）组织稀缺——可能仅有一块组织切片可用于染色；（2）成本限制——IHC 染色的成本显著高于 H&E；（3）工作流程异步——H&E 和 IHC 通常在不同时间处理，可能来自不同的组织块或活检部位。

ACM MM'25 论文通过双分支架构解决了这一问题，该架构根据轻量级模态分类器动态选择交叉模态重建路径（用于单模态输入）或端到端融合管道（用于双模态输入）。这种设计使得框架能够在无论哪种模态可用的情况下保持高准确率，而无需为训练提供昂贵的配对数据集。

2.3 来源论文的范围和贡献

来源论文做出了四个核心贡献：（1）基于 Pyramid Pix2Pix 的交叉模态生成对抗网络（CM-GAN），用于特征空间中的双向 H&E↔IHC 重建；（2）使用领域分类损失（DCO）和分布对齐损失（DAO）的病理信息解耦编码器，以分离共享和模态特定特征；（3）一个受 CBAM 启发的模态敏感特征注意模块，用于自适应模态加权；（4）在 BCI（乳腺癌免疫组化）数据集上进行全面评估，包含 4,870 对注册的 H&E-IHC 图像。

3. 来源材料的实证发现

3.1 双模态灵活架构

该框架依次采用四个核心模块。首先，一个 **分支选择器**（99.95% 分类准确率）确定输入是单模态还是双模态，并相应路由。其次，对于单模态输入，**CM-GAN** 重建缺失的模态——该模块使用一个双向 GAN，可以从 H&E 输入合成 IHC 特征，反之亦然，基于 BCI Pyramid Pix2Pix 架构进行构建并进行了修改。第三，对于单模态和双模态输入，**病理信息解耦编码器**（预训练的 ResNet50 主干）分离共享特征（H&E 和 IHC 共有的形态模式）与模态特定特征（H&E 捕捉组织结构；IHC 捕捉蛋白表达模式），使用 DCO 将共享和特定特征推开，使用 DAO 对齐跨域特征分布。第四，**受 CBAM 启发的模态敏感注意模块**在应用通道注意之前连接共享和特定特征向量，根据输入质量自适应地重新加权模态贡献。

3.2 性能结果

论文中报告的关键性能结果为：双模态（真实 H&E + 真实 IHC）实现 95.09% 的准确率（F1 = 0.9532）；跨模态（真实 H&E + 虚假 IHC）实现 94.25% 的准确率（F1 = 0.9609），比 H&E 单模态基线提高了 +22.81%，跨模态（真实 IHC + 虚假 H&E）实现

90.28% 的准确率（F1 = 0.9038），比 IHC 单模态基线提高了 +12.90%；注意模块在无注意基线基础上贡献了额外的 +4.91% F1 改进。

3.3 重建质量

跨模态重建质量指标：H&E 到 IHC 的重建实现 PSNR = 18.48 dB，SSIM = 0.51；IHC 到 H&E 的重建实现 PSNR = 17.24 dB，SSIM = 0.39。这些 SSIM 值在自然图像合成文献中相对较低——作为背景，BCI Pyramid Pix2Pix 基线（P012）在同一任务上实现 SSIM=0.431，相较之下 CM-GAN 的 0.51 代表了大约 18% 的相对 SSIM 改进，但两者仍远低于通常被认为是高质量的 SSIM > 0.90 范围。

3.4 识别的限制

论文自身识别的限制包括：单一机构评估（BCI 数据集，单一机构）、模型输出中没有不确定性量化、没有每类准确率报告（特别是 HER2 1+ 缺失）、没有与最先进的 HER2 AI 系统进行比较、类不平衡未量化、没有孤立 DCO 与 DAO 贡献的消融实验。

3.5 数据集和架构细节

BCI 数据集包含来自单一机构的 4,870 对注册的 H&E-IHC 图像，涵盖所有四个 HER2 表达水平。ResNet50 主干在 ImageNet 上预训练，并使用混合解冻策略进行微调（第 1–2 层冻结以提取通用特征，第 3–4 层 + fc 微调）。CM-GAN 使用跨四个高斯金字塔层的多尺度残差块，结合 GAN 损失和 L1 像素损失进行训练。

4. 基于文献的深度分析

4.1 保留的详细论文分析

论文 1: 可解释且鲁棒的深度学习用于自动化 HER2 评估 (P001, Research Square, 2026)

问题和任务设定: 该论文解决了自动化 HER2 IHC 评分中的三个相互关联的挑战：深度学习模型的有限可解释性（黑箱问题）、不同放大倍数下的鲁棒性差（10× 与 40×）、与病理推理的对齐不足。任务是对三个公共数据集进行 4 类 HER2 表达分类（0, 1+, 2+, 3+）：BCI（10×）、HER2-IHC-40x-Patch 和 HER2-IHC-40x-WSI。

方法论: 该框架采用 ResNet50，比较了三种训练配置：冻结主干（仅训练 fc 层）、完全微调（训练所有层）和混合解冻（第 1–2 层冻结以提取通用特征，第 3–4 层 + fc 微调）。混合方法的动机在于早期 ResNet 层捕捉通用组织结构（细胞膜、细胞核、组织模式），而后续层编码放大倍数特定和任务特定的特征。为了可解释性，Score-CAM 生成注意图，并引入了两个量化指标：MAP（膜激活精度），测量模型注意与病理学家注释的膜区域之间的重叠，以及 EC（解释一致性），测量相似输入之间解释的稳定性。评估框架还包括期望校准误差（ECE），以评估模型置信度的可靠性。训练使用 Adam 优化器（lr=0.001）、余弦退火调度器、早停（耐心=5）、L2 正则化（1×10^{-4}）和数据增强（水平翻转、$\pm20^{\circ}$ 旋转、颜色抖动）。批量大小为 32，最大 100 个周期。

主要证据: 混合解冻配置在 BCI 上实现 95% 的准确率（与 73% 的冻结主干、91% 的完全微调相比），在 HER2-IHC-40x-Patch 上实现 96%（与 91% 的冻结主干、95% 的完全微调相比）。混合解冻的改进在 BCI 的 10× 放大倍数上最为明显，完全微调仅实现 91%，而混合解冻实现 95%。与完全微调相比，计算开销减少了 72.7%。每类 BCI 结果：类 0（P=0.97，R=0.97），类 1+（P=0.95，R=0.90），类 2+（P=0.94，R=0.98），类 3+（P=0.97，R=0.94）。在 HER2-IHC-40x-WSI 上：类 1+（P=0.82，R=0.83）——显著低于 BCI，表明对数据集分布的敏感性。MAP 分数为 79%，表明模型的 79% 注意与病理学家注释的膜区域重叠。HER2 3+ 鉴别的 AUC > 0.99。

相关性: 混合解冻方法在架构上与 MM'25 论文的预训练 ResNet50 主干互补——MM'25 论文可以从层冻结策略中受益，以降低计算成本，同时保持性能。MAP/EC 指标为评估 MM'25 论文的注意模块是否产生临床有意义的解释提供了具体的评估框架。BCI 数据集在 10× 的结果直接展示了放大倍数鲁棒性挑战：冻结主干在 BCI 上降至 73%，表明 MM'25 论文在 BCI 补丁上的单一放大倍数评估可能无法充分表征跨放大倍数的泛化。

限制: 专注于 IHC 图像（而非 H&E+IHC 多模态），且未解决缺失模态场景。类 1+ 的 F1=0.85 在 BCI 10× 上反映了 HER2-low 边界的已知困难。未进行正式的病理学家验证 MAP/EC 指标——仅与病理注释的重叠进行了自动化比较。数据集来源为单一机构（BCI），与 MM'25 论文相同。

论文 2: 使用多模态 MRI 进行 HER2 表达预测 (P002, Frontiers in Oncology, 2025)

问题和任务设定: 解决从 MRI 成像（T1、T2、增强对比序列）结合临床列线图特征进行 HER2 预测的问题，适用于 IHC 结果不可用或不确定的患者。这是一种与组织病理学根本不同的成像模态。

方法论: 深度学习（ResNet50、VGG16、EfficientNet-B0、ViT-Small）结合临床列线图特征，使用后期融合。使用 ICC 过滤和 LASSO 回归进行特征选择。AUC-ROC 作为主要指标。三类分类（HER2 阳性、HER2 阴性、可疑）。后期融合方法涉及独立训练每种成像模态主干，从最终的全连接层提取特征，将其与临床列线图特征（肿瘤大小、分级、Ki-67 指数等）连接，并通过分类头传递组合特征向量。

主要证据: 使用 ResNet50 + 临床数据集成的 AUC = 0.94。后期融合方法（ResNet50 + 临床列线图）实现了最高的 AUC，优于单一模态主干和仅成像的融合。证明了从非 IHC 成像模态进行 HER2 预测是可行的，并且将结构成像特征与临床元数据结合提供了互补的诊断信息。

相关性: 证实了 MM'25 论文的前提，即缺失模态的 HER2 预测在临床上具有价值。后期融合架构（P002）代表了 MM'25 论文共享特定解耦方法的替代方案：P002 在决策层面连接来自不同模态的特征，而 MM'25 论文在融合之前将特征分解为共享和特定流。P002 的发现表明临床列线图特征（肿瘤分级、Ki-67）在成像之外增加了判别价值，暗示 MM'25 框架的潜在扩展——结合临床元数据与组织病理学特征可能改善 HER2 分类的鲁棒性。不同的成像模态

（MRI 与组织病理学）提供互补的生物信息，尽管 MM'25 论文的具体贡献仍然是 H&E+IHC 组织病理学多模态融合方法。

限制: 与 MM'25 论文不同的成像模态；基于 MRI 的 HER2 预测尚未在此特定任务上进行临床验证。后期融合方法在架构上与 MM'25 论文的特征解耦编码器不同，且尚未在相同的 HER2 分类任务上进行验证。

论文 3: HER2-low 乳腺癌 — 识别、检测和治疗中的挑战 (P003, PMC, 2024)

问题和任务设定: HER2-low（IHC 1+ 或 2+ 且 ISH 阴性）占乳腺癌的 ~40-50%，并且可以用 T-DXd 进行治疗，因此准确识别至关重要。挑战在于 HER2-low 与 HER2 0 之间存在连续性，使得 0/1+ 边界成为诊断上最困难的。

方法论: 系统回顾了在前分析（组织固定、处理）、分析（抗体克隆、评分指南）和后分析（解释者变异性）因素中的挑战。肿瘤内异质性被识别为 HER2 状态变异的主要原因。该回顾综合了临床研究、病理指南和新兴计算病理学文献中的证据。

主要证据: 方法变异的最大影响发生在 IHC 0-1+ 接口。HER2-low 状态的动态特性：HER2 表达在全身治疗后可能发生变化，需要重新评估生物标志物。对于可疑（2+）病例，观察者间一致性在 0/1+ 边界报告为 60-75%，而边界病例的结果则显著更低。计算病理学被确定为一个有前景的方向，但尚未在临床上验证。回顾指出，HER2 评分中的观察者间变异性是一个众所周知的问题，在 HER2-low 边界处的分歧最大。

相关性: 为 MM'25 论文的动机提供了临床背景，并量化了诊断挑战的规模。0/1+ 边界是最可变的区域——根据系统回顾，对于 2+ 病例的观察者间一致性为 60-75%——与 MM'25 论文的评估直接相关。MM'25 论文未报告每类准确率，未明确框架的 94.25% 总体准确率是否掩盖了 HER2 1+ 病例的性能下降。准确检测 HER2-low 的临床迫切性（T-DXd 的治疗资格）加强了填补这一空白的紧迫性。HER2 表达在全身治疗后可能发生变化的发现也提出了一个纵向验证的考虑，而 MM'25 论文的横断面评估未能解决这一问题。

限制: 综述文章；未提出计算解决方案。引用的具体性能数字（60-75% 一致性）是针对可疑 2+ 病例，而非 0/1+ 边界。

论文 4: 计算病理学在 HER2-low 识别中的应用 (P004, PMC, 2024)

问题和任务设定: 确定计算病理学在 HER2-low 识别中的机会和挑战，重点分析研究原型与可临床部署系统之间的差距。分析已发表的 AI 方法以评估该领域的临床部署准备情况。

方法论: 分析已发表的 HER2 评分 AI 方法，识别三个主要差距：（1）缺乏针对 HER2-low 的标准化评估协议；（2）训练数据多样性有限；（3）模型输出中缺乏不确定性量化。该论文调查了包括深度学习分类器、基于注意力的模型和基础模型方法的研究，评估它们在 HER2-low 案例上的表现。

主要证据: 没有单一的已发表的计算病理学方法在常规临床实践中得到验证以进行 HER2-low 分类。染色变异、组织异质性和阅读者变异性被识别为临床部署的三大障碍。文献回顾发现，大多数已发表的 HER2 AI 研究报告总体准确率，但未能提供每类性能的细分，特别是 HER2 1+。回顾还指出，现有的 HER2 AI 系统尚未在临床环境中进行前瞻性验证。

相关性: 识别的三个差距直接映射到 MM'25 论文的限制：单一机构 BCI 数据集（数据多样性差距）、未报告每类性能（未进行 HER2-low 特定评估）和缺乏不确定性量化（临床部署差距）。发现 HER2-low 识别是整个领域未解决的挑战——不仅仅是 MM'25 论文——为论文的贡献提供了背景：它解决了更广泛社区尚未解决的重要问题。该论文识别染色变异作为最大部署障碍的发现直接证实了 MM'25 论文评估提出的单一机构有效性问题。

限制: 没有新的实验结果；依赖于文献综合。具体障碍量化（染色变异作为“最大”）是定性的而非定量的。

论文 5: ResViT-GANNet — 基于多模态注意力和 GAN 的乳腺癌组织病理学增强 (P005, BMC Medical Imaging, 2025)

问题和任务设定: 使用结合 CNN 和视觉变换器（ViT）、令牌对齐多模态注意（TAMA）和 StyleGAN2-ADA 的多模态架构对乳腺癌组织病理学图像进行分类。已在覆盖 8 种肿瘤类型和多个放大倍数的多机构 BreakHis 数据集上进行验证。

方法论: 双分支架构：CNN 分支（ResNet50/EfficientNet 主干）用于 H&E 图像 + ViT 分支用于额外成像数据。TAMA 模块使用 CNN 和 ViT 令牌序列之间的交叉注意对来自两个分支的异构特征进行对齐和融合。StyleGAN2-ADA 在训练期间生成用于数据增强的合成组织病理学图像。该框架在 8 种放大倍数（40×、100×、200×、400×）和 8 种肿瘤类型（乳腺癌亚型）上处理图像。

主要证据: 在乳腺癌组织病理学分类中实现 96.40% 的准确率，合成数据的引入提高了 3.3%。TAMA 模块在多模态设置中优于后期融合，在 BreakHis 上实现了 2.1% 的更高准确率。消融研究表明，GAN 增强在少数肿瘤类别上贡献了 3.3% 的准确率提升。重要的是，ResViT-GANNet 在 BreakHis 上得到了验证，这是一个具有固有染色变异的多机构数据集——使得 96.40% 的准确率比单一机构结果更具鲁棒性。

相关性: 与 MM'25 论文共享双分支多模态架构概念，并在多机构数据上验证了这一架构家族。TAMA 模块在架构上与 MM'25 论文的 CBAM 启发的模态敏感注意相似——两者都使用交叉注意机制动态加权多模态特征。96.40% 的准确率在多机构 BreakHis 上尤其重要：这表明在现实世界的多机构条件下可以实现 ~96% 的准确率，暗示 MM'25 论文在单一机构 BCI 上的 95.09% 双模态准确率在部署中是一个合理的性能范围。GAN 增强的发现（3.3% 的改进）与 MM'25 论文的 CM-GAN 跨模态合成概念相关，表明生成模型在组织病理学分类中具有可测量的价值。然而，ResViT-GANNet 并未解决缺失模态场景——它假设两种模态均可用——因此无法直接替代 CM-GAN 的缺失模态鲁棒性贡献。

限制: 不解决缺失模态场景；未在 BCI 数据集或 HER2 特定任务上进行评估；合成增强和跨模态重建是概念上相关但不同的任务。BreakHis 数据集使用不同的肿瘤类型而非 HER2 表达水平，限制了与 MM'25 论文任务的直接可比性。

论文 6: MoRA — 基于 LoRA 的多模态疾病诊断与缺失模态 (P006, arXiv, 2024)

问题和任务设定: 多模态预训练模型在缺失模态时性能急剧下降，且完全微调计算成本高。MoRA 通过在预训练多模态模型的第一个变换器块上使用模态感知的低秩适应来解决这两个挑战。

方法论: MoRA 将每个输入投影到低内在维度，并使用模态感知的上投影。MoRA 为不同的模态可用性场景（完整、缺失模态 1、缺失模态 2）使用不同的上投影矩阵，而不是为所有模态计算共享的 LoRA 投影。当模态 m1 缺失时，选择该场景的相应投影。关键创新在于仅 1.6% 的参数可训练，而不是完全微调。发现 r=4 在 ODIR 上是最佳的。该方法集成到第一个变换器块中，因为第一层可以直接获取输入令牌的信息，有助于确认缺失模态的状态。数据集：CXR（3,030 训练，385 验证，379 测试，20 种疾病）和 ODIR（2,781 训练，382 验证，337 测试，7 种疾病）。

主要证据: 在 CXR 上，图像缺失率为 30% + 文本 100% 时，MoRA 实现 F1-Macro=37.22，MAPs=33.49，MSPs=35.13，ViLT=25.54。在 ODIR 上，图像缺失率为 30% + 文本 100% 时，MoRA 实现 F1-Macro=92.56，MAPs=90.66，MSPs=46.38，ViLT=81.34。当文本是 ODIR 上缺失的模态（100% 图像，30% 文本）时，MoRA 实现 F1-Macro=76.89，MAPs=78.71——在这种情况下，MAPs 略微优于 MoRA。GPU 内存：MoRA 在 CXR 上需要 12.2 GB，而 MAPs 需要 14.4 GB；训练时间为 1.58 小时，而 MAPs 为 1.71 小时每 1,000 步。秩敏感性：r=4 是最佳的（F1=80.73）；r=384（全秩）给出 70.23 F1-Macro，低于没有 MoRA 的情况，确认低秩是必需的。

相关性: 提供了缺失模态鲁棒性的最强替代方法，而无需生成图像合成。模态感知的上投影机制在概念上与 MM'25 论文的分支选择器 + CM-GAN 管道相关：两者都根据模态可用性选择不同的处理路径。然而，MoRA 需要预训练的多模态变换器，而 MM'25 论文从头开始训练。在相同的 HER2 分类任务上对 CM-GAN 与 MoRA 进行比较基准测试将澄清 CM-GAN 的生成方法是否优越。MoRA 的参数效率（1.6%）与 CM-GAN 的完整生成管道形成对比——如果 MoRA 在 HER2 分类上实现可比准确率，则将代表一种更高效的解决方案。

限制: 需要预训练的多模态模型；不直接适用于 MM'25 论文的从头训练设置；没有针对组织病理学的特定评估。该方法在图像+文本（CXR、ODIR）上进行评估，而不是图像+图像（H&E+IHC），因此直接转移到 MM'25 论文领域的有效性尚未得到证明。

论文 7: SimMLM — 具有缺失模态的多模态学习 (P007, ICCV 2025)

问题和任务设定: 现有的缺失模态方法依赖于复杂的架构或数据插补。SimMLM 提供了一个通用框架，具有动态模态专家混合（DMoME）和新颖的更多与更少（MoFe）排名损失，确保准确性单调性：添加模态应提高或保持准确性。

方法论: DMoME 由模态特定的专家网络（每个都是独立的主干）和一个可学习的门控网络组成，该网络在 logit 级别动态调整每个模态的贡献（在 softmax 之前）。MoFe 损失强制执行 L*MoFe = max(0, L*task(o*fewer, y) - L*task(o*more, y))，其中 o*more 的模态比 o_fewer 多。两阶段训练：（1）独立预训练专家，最小化干扰；（2）合作学习，联合训练专家 + 门控 + MoFe 损失。logit 级别的加权为模型校准提供了温度缩放效果。在 BraTS 2018（脑肿瘤分割，4 种 MRI 模态，15 种模态配置）、UPMC Food-101（图像+文本）和 avMNIST（图像+音频）上进行了验证。

主要证据: 在 BraTS 2018（15 种模态配置）上，SimMLM 在所有设置中实现了最高的平均 Dice 分数。在 UPMC Food-101 上，SimMLM 实现了 72.20% 的图像仅（vs. MoMKE 70.46%）、87.2% 的文本仅（vs. MoMKE 86.59%）和 94.99% 的全模态（vs. MoMKE 92.71%）。在 avMNIST 上，94.27% 的全模态。校准（ECE↓）：SimMLM 实现了 3.15%/3.75%/3.55%（ET/TC/WT）vs. MoMKE 3.46%/4.10%/4.04%。MoFe 损失的最佳系数 λ=0.1。专家预训练至关重要：跳过会导致显著的性能下降。logit 级别的混合优于特征级和概率级的混合。

相关性: DMoME 门控机制在架构上与 MM'25 论文的 CBAM 启发的通道注意相似，后者在通道加权之前连接共享和特定特征——两者都自适应地加权模态贡献。MoFe 排名损失可以应用于 MM'25 框架的训练，以确保双模态准确性 ≥ 单模态准确性，提供一个原则性的训练约束。SimMLM 的校准结果（ECE 改进 0.31–0.35 个百分点）与 MM'25 论文中识别的临床部署差距特别相关——不确定性量化是 HER2 AI 的可信赖性所需。

限制: 在非组织病理学任务（脑肿瘤分割、食品分类、数字分类）上进行验证。直接适用于 HER2 预测在组织病理学图像上的有效性尚未得到证明。MoFe 损失需要具有不同模态可用性的训练数据，这可能与配对 H&E+IHC 训练设置不一致。

论文 8: MODES — 临床诊断的解耦多模态表示融合 (P008, npj Digital Medicine, 2025)

问题和任务设定: 简单的多模态连接会导致语义干扰和信息丢失。MODES 明确解耦共享特征（跨模态的共同特征）与模态特定特征，以提高融合质量。在使用来自 UK Biobank 的 ECG 和心脏 MRI（cMRI）的心血管模型上进行了评估（4,150 个持出样本）。

方法论: 四个组件：（1）对共享（Z^s）和模态特定（Z^n）表示进行编码的单模态预训练编码器（基础模型）；（2）重建原始数据的单模态生成器；（3）具有掩蔽组件的潜在表示，消除低信息维度；（4）迭代三步训练：编码共享+特定 → 掩蔽优化 → 生成器重建。掩蔽组件使

用二进制掩蔽，L1 正则化由 β 控制，使模型能够推断适当的表示维度。比较基线：单模态表示、早期融合、后期融合、DRIM（解耦基线）。

主要证据: MODES 表示在一系列诊断表型（RR 间隔、射血分数等）和诊断（房颤、瓣膜疾病）上优于所有基线。缺失模态：ECG 表示可以推断与 cMRI 相关的表型，反之亦然，通过共享表示。掩蔽验证：使用掩蔽的预测性能几乎与不使用掩蔽相同，确认紧凑表示而不丧失信息。共享表示捕捉跨模态的诊断信息，而模态特定表示编码独特信息（ECG 特定：电活动；cMRI 特定：心脏解剖）。掩蔽在不同模态对之间收敛到不同的子空间大小，具体取决于信息内容。

相关性: 为 MM'25 论文的病理信息解耦编码器提供了最强的方法论验证。MODES 在完全不同的领域验证了共享特定解耦的原则，确立了它作为一种领域无关的原则。MM'25 论文中的 DCO（领域分类损失）和 DAO（分布对齐损失）来自 Wang 等人的 CVPR 2023，代表了在组织病理学中类似 MODES 的解耦的领域特定实例。掩蔽可以在不损失准确性的情况下减少表示维度的关键见解适用于 MM'25 论文的注意机制设计。

限制: 在心脏成像中进行评估，而非组织病理学。MM'25 论文的基于 ResNet50 的共享编码器与 DCO/DAO 损失是领域特定的实例。MODES 需要预训练的基础模型，而 MM'25 论文没有使用。

论文 9: CPath-Omni — 计算病理学的统一多模态基础模型 (P009, arXiv, 2024)

问题和任务设定: 现有的计算病理学模型处理的是补丁级或全切片图像分析，而不是两者兼顾。CPath-Omni 是一个 150 亿参数的多模态 LMM，支持这两项任务，将病理图像与临床文本相结合，使用自监督目标进行训练。

方法论: 四阶段训练：（1）补丁级预训练，将 CPath-CLIP 特征与 Qwen2.5-14B LLM 对齐，使用 700,145 对图像-标题对（CPath-PatchCaption）；（2）使用 CPath-PatchInstruct 进行微调，进行 VQA、分类、标题生成；（3）使用 CPath-WSIInstruct 进行 WSI 预训练，使用 SlideParser 进行多尺度 WSI 令牌化；（4）混合补丁-WSI 训练（15% 随机抽样）以进行知识转移。CPath-CLIP 结合 OpenAI-CLIP-L 和 Virchow2（基于 DINOv2，3M WSI）作为双视觉编码器，Qwen2.5-14B 作为文本编码器。SlideParser 执行多尺度区域编码（10×/20×/40×）和令牌压缩（CoCa 风格的 1,152 个查询令牌）以标准化可变大小的 WSI 输入。

主要证据: CPath-Omni 在 39/42 个基准测试中实现了 SOTA。在 PathMMU（最大的病理 VQA 数据集）上，CPath-Omni 超过了 PathGen-LLaVA 13.8%，并超越了人类病理学家的表现（71.8%）0.6%。在 CPath-CLIP 零-shot 分类中，CPath-Omni 在 Osteo/Pcam/LC-Lung 上超越了 PathGen-CLIP-L 6.1%/7.7%/7.3%。少量样本：在仅有 2 个样本的情况下，CRC 的准确率为 95%（vs. 其他模型 <91%）。在 WSI 级任务中，CPath-Omni 的表现与任务特定的 ABMIL/DSMIL 模型相当。训练仅使用 700K 图像-标题对，远少于通用领域 CLIP 模型。

相关性: 代表了计算病理学基础模型的前沿。CPath-Omni 的双视觉编码器设计（CLIP + DINOv2）及其多尺度 WSI 处理在架构上与 MM'25 论文的 CNN 基础方法不同，但统一模型能够匹配任务特定模型的发现对多模态 HER2 AI 的长期前景是令人鼓舞的。CPath-Omni 没有解决缺失模态鲁棒性，这表明即使在基础模型规模上，这一特定能力仍然是一个开放的研究问题——验证了 MM'25 论文的具体贡献。

限制: 需要巨大的计算资源（150 亿参数）；未具体解决缺失模态处理；未报告 HER2 特定评估；从头训练与 MM'25 论文的微调场景不同。

论文 10: 使用多中心数据集进行染色标准化基准测试 (P010, Scientific Reports, 2026)

问题和任务设定: H&E 染色的组织标本在不同机构、扫描仪和染色协议之间表现出显著的颜色和染色变异。本研究在一个独特的多中心数据集上基准测试了八种染色标准化方法：相同的组织块（结肠、肾脏、皮肤）在 66 个不同的实验室染色，隔离染色变异与其他生物/技术变异。

方法论: 四种传统方法（直方图匹配、Macenko、Vahadane、Reinhard）和四种深度学习方法（CycleGAN-UNet、CycleGAN-ResNet、Pix2pix-UNet、Pix2pix-DenseUNet）。定量评估：颜色转移指标（交集、PCC、欧几里得距离、lαβ 空间中的 JS 散度）、SSIM 结构相似性、高级特征相似性的 FID、Cellpose-SAM 细胞核检测计数，以及基础模型（UNI-2）特征提取与 t-SNE 可视化。WSI 重采样为 10× 以提高计算效率。对传统和深度学习方法应用整体 WSI 标准化（而非补丁级）以避免平铺伪影。

主要证据: 没有单一方法在所有组织类型和所有指标上占主导地位。颜色转移：直方图匹配在皮肤（交集=0.891，PCC=0.938，FID=61.67）和肾脏（交集=0.944，PCC=0.985）上实现最佳平均分数，结肠：Macenko 和 Reinhard 表现相当。结构相似性：Vahadane 实现最高 SSIM（0.995），但颜色转移最差；所有方法的 SSIM 均保持在 > 0.92。细胞核检测：Pix2pix-UNet/CycleGAN-ResNet/CycleGAN-UNet yield 计数最接近参考；在传统方法中，直方图匹配最佳。基础模型特征（t-SNE）：CycleGAN 和 Macenko 产生最紧凑的聚类；Vahadane 和 Pix2pix-DenseUNet 产生分散的分布。推理时间：CycleGAN/Pix2pix ~4–5 分钟/WSI；直方图匹配/Reinhard ~30 秒–2 分钟/WSI；Macenko/Vahadane ~2–7 分钟/WSI。深度学习方法产生幻觉伪影（CycleGAN-ResNet：在脂肪组织中产生假细胞核；Pix2pix-DenseUNet：产生错误着色的平滑肌细胞核）。传统方法：Macenko 在伊红中产生蓝色伪影；Vahadane 在许多情况下完全丢弃苏木精。

相关性: 直接验证了 MM'25 论文 BCI 数据集评估的单一机构限制。没有单一染色标准化方法占主导地位的发现表明，CM-GAN 需要在评估鲁棒性之前将染色标准化作为预处理步骤。深度学习基于标准化方法（CycleGAN、Pix2pix）产生的幻觉伪影与评估 CM-GAN 的跨模态 IHC 生成伪影是否可能损害下游分类直接相关。66 个实验室数据集可作为未来跨机构 HER2 AI 验证的基准。

限制: 在 H&E 染色上进行评估，而非 IHC。MM'25 论文的 CM-GAN 从 H&E 生成 IHC，这与 H&E 到标准化 H&E 的标准化任务不同。IHC 特定的染色变异可能在幅度和特征上与 H&E 变异不同。

论文 11: 针对组织切片的多目标染色标准化 (P011, arXiv, 2024)

问题和任务设定: 标准染色标准化使用单一参考图像，这未能捕捉到实际数据集中染色模式的多样性。多目标标准化使用多个参考图像，以提高对多机构环境中染色模式多样性的鲁棒性。

方法论: 无参数方法，使用多个参考图像针对每种目标染色。该方法不是从单一典型参考图像计算染色统计，而是从一组多样化的代表性图像学习染色统计，这些图像捕捉到实践中观察到的染色变异范围。该方法适应目标数据集中染色颜色的统计分布，而不是强迫所有图像朝向单一参考调色板。

主要证据: 与单一参考方法相比，提高了对染色变异的鲁棒性，更好地推广到外部数据集。具体而言，多参考方法避免了对单一参考图像的特定颜色分布过拟合的陷阱，这是单一参考方法在外部数据上部署时的主要失败模式。该方法不需要染色颜色参数的调优——它直接从参考集学习适当的标准化目标。

相关性: 直接解决了 P010 提出的跨机构泛化问题，即没有单一染色标准化方法占主导地位。MM'25 论文依赖于 CM-GAN 合成缺失模态，但未对现有输入模态进行标准化——多目标标准化可以作为 CM-GAN 之前的输入预处理步骤，可能减少来自不同机构的 H&E 图像之间的分布偏移。该方法对 MM'25 论文特别有前景，因为它可能减少 CM-GAN 的跨模态重建质量在染色变异下的方差，解决 SSIM-准确率解离的伪影利用假设（H1）。与单一参考方法相比，多目标标准化本质上对领域转移更具鲁棒性，因为它不承诺于单一染色颜色分布。

限制: arXiv 预印本；有限的实证验证细节。可用片段中未报告针对单一参考方法的定量比较。未报告的具体定量改进（例如，“在外部数据上比 Macenko 的颜色转移好 X%”）在可用片段中未列出。

论文 12: 通过 Pyramid Pix2pix 生成乳腺癌免疫组化图像 (P012, CVPR 2022)

问题和任务设定: H&E 到 IHC 跨模态翻译问题的基础数据集和基线方法。BCI 数据集包含 4,870 对注册的 H&E-IHC 图像，涵盖所有四个 HER2 表达水平（0、1+、2+、3+）。任务是从 H&E 图像合成 IHC 图像，以减少实际 IHC 染色的成本和延迟。

方法论: Pyramid Pix2pix 架构：跨四个高斯金字塔层的多尺度残差块，具有跳跃连接。多尺度损失：L*multi-scale* = $\Sigma_i \lambda_i S_i$，*结合跨尺度的 GAN 损失（L*cGAN）和 L1 像素损失。数据集构建：未染色的组织 → HE 染色 → IHC 染色 → elastix 注册（为提高效率进行 16 块并行化）→ 图像精细化 → 切割。通过与投影变换的重叠比较验证注册质量。

主要证据: Pyramid Pix2pix 在 BCI 基准上优于标准的 Pix2pix、Pix2PixHD 和 CycleGAN。伽玛校正变体在同一任务上实现 PSNR=16.024 dB，SSIM=0.431（加权 0.6×SSIM + 0.4×PSNR 排名）。Pix2pixHD 在低 HER2 表达区域错误生成深棕色。Pyramid Pix2pix 在图像质量和 HER2 表达识别方面均优于 CycleGAN 和 Pix2pix 变

体。病理学家评估：两位病理学家对生成的 IHC 图像的准确率为 37.5% 和 40.0%——确认生成的图像在图像级别尚未具备临床可解释性。该方法在低 HER2 表达（0/1+）方面表现优于高表达（3+）的生成图像真实性。

相关性: MM'25 论文的 CM-GAN 直接基于 BCI 的 Pyramid Pix2Pix 作为其跨模态模块的基础。MM'25 论文的 CM-GAN 实现 PSNR=18.48 dB，SSIM=0.51，代表了相对于 BCI 伽玛校正变体（SSIM=0.431）约 18% 的相对 SSIM 改进，以及 PSNR 的 15% 改进。然而，两者的 SSIM 值仍然相对较低。关键是，MM'25 论文的下游分类准确率（94.25%，真实 H&E + 虚假 IHC）表明分类头能够容忍相当大的重建噪声——问题在于这种容忍是否能推广到跨机构的部署。病理学家评估确认生成的 IHC 图像在临床上无法被人类专家解释，这限制了生成图像的临床实用性。37.5–40% 的病理学家准确率提供了 CM-GAN 在人类可解释性方面应测量的具体基线。

限制: 仅在图像合成质量指标上进行评估，而未在合成图像的下游 HER2 分类准确率上进行评估。BCI 数据集来自单一机构，注册过程（16 块 elastix）可能在配对数据集中引入伪影。

PDF 精炼增强分析

下载的 10 篇论文的 PDF 提供了大量额外的定量细节，增强了上述分析：

- **P001（HER2 可解释评估）**: PDF 提供了完整的每类分类表（表 5），显示 HER2 2+ 是 HER2-IHC-40x-WSI 上最具挑战性的类别（P=0.80，R=0.97），而 HER2 3+ 在 HER2-IHC-40x-Patch 上实现完美分类。MAP=79% 的分数在结论中量化。混合解冻消融（表 9）量化了冻结主干在 BCI 上的严重降级（准确率 73%，宏 F1=0.69）与所提方法（准确率 95%，宏 F1=0.95）之间的具体证据。
- **P006（MoRA）**: PDF 提供了完整的表 2-5，包含不同模态缺失配置下的所有 F1-Macro 数字。秩敏感性分析（表 5）确认 r=4 是最佳的（F1=80.73），而 r=384（全秩）给出 70.23 F1-Macro，低于没有 MoRA 的情况，严格验证了低秩假设。GPU 内存和训练时间数据（表 3）量化了计算优势。
- **P007（SimMLM）**: PDF 提供了完整的 BraTS 2018 基准表（表 1，包含 15 种模态配置）、UPMC Food-101/avMNIST 准确率表（表 2）和校准误差表（表 4）。logit 级别与特征级和概率级混合的消融（图 A5）与具体 ECE 数字确认了 logit 级别设计选择。MoFe 消融与 λ 敏感性数据提供了最佳 λ=0.1 的发现，并有定量支持。
- **P008（MODES）**: PDF 提供了四个组件框架的详细信息（编码器、生成器、表示、掩蔽）、三步迭代训练程序和定量表型预测结果。共享与特定表示在 bMRI-cMRI 模态对上的分析确认某些模态对几乎没有共享信息（例如，ECG 的电气指标与 cMRI 的心脏解剖），验证了不同领域中的解耦方法。
- **P009（CPath-Omni）**: PDF 提供了完整的四阶段训练程序、CPath-CLIP 架构细节（双视觉编码器组合）和 42 个数据集的定量结果。人类水平的性能比较（专家 72.3% vs. CPath-Omni 71.8% 在 PathMMU 上）提供了 0.6% 的盈余数字。

- **P010（染色标准化）**：PDF 提供了所有八种方法在三种组织类型上的完整定量结果、幻觉伪影描述和基础模型特征分析与 t-SNE。推理时间比较量化了方法之间的计算权衡。引用的具体值：直方图匹配在皮肤上的交集=0.891；CycleGAN FID=61.67；Vahadane SSIM=0.995。
- **P012（BCI）**：PDF 提供了 Pyramid Pix2pix 架构图（跨高斯卷积的多尺度金字塔）、多尺度损失公式、使用 16 块并行化的 elastix 注册管道和病理学家准确率评估（37.5%/40.0%）。数据集统计（3,123 WSI 对，1,750 补丁对）和类分布（图 8）被提取。

4.2 综合主题评估

4.2.1 主题：缺失模态问题及其对 HER2 临床部署的重要性

缺失模态鲁棒性在 HER2 预测中的临床动机令人信服，并得到了文献的充分支持。系统回顾（P003）记录了 HER2-low（~40-50% 的乳腺癌）是治疗上最重要且诊断上最困难的类别，并且在 0/1+ 边界处的观察者间变异性（2+ 病例的 60-75% 一致性）是人工评分的根本限制。存在一种自动化方法能够仅从 H&E 预测 HER2 表达（94.25% 的准确率，较基线提高 +22.81%），将改变资源有限地区的临床工作流程。计算病理学回顾（P004）确认，现有的 HER2 AI 系统尚未在常规临床实践中验证 HER2-low 的识别，突显了临床紧迫性和该领域当前的空白。

文献揭示了三种不同的架构家族来处理多模态深度学习中的缺失模态鲁棒性，每种都有不同的权衡，相关于 MM'25 论文的方法。第一类是 **生成图像合成**，以 CM-GAN 及其 BCI Pyramid Pix2Pix 基础（P012）为例。这种方法的优势在于它生成可视化的合成图像——即使 SSIM 较低，病理学家原则上可以检查生成的模态。关键限制是 BCI 病理学家在解释生成的 IHC 图像时仅获得 37.5–40% 的准确率（P012），确认低 SSIM 合成生成的图像会混淆甚至专家观察者。第二类是 **模态感知的参数高效适应**，以 MoRA（P006）为例，该方法使用模态特定的低秩投影，仅需 1.6% 的可训练参数。这种方法在计算上比生成合成高效得多，并在图像+文本模态上取得了良好的结果，但需要预训练的多模态变换器，并且尚未在组织病理学图像+图像任务上进行验证。第三类是 **动态专家门控**，以 SimMLM 的 DMoME（P007）为例，使用 logit 级别的模态专家混合和强制准确性单调性的排名损失。这种方法在架构上与 MM'25 论文的 CBAM 启发的注意模块最为接近，其校准改进（在 BraTS 上 ECE 降低 0.31–0.35 个百分点）与 MM'25 论文中识别的不确定性量化差距直接相关。

选择生成合成而非特征/令牌级适应的结构动机根植于跨模态翻译任务的性质。H&E 到 IHC 的翻译涉及像素级的语义映射——将 H&E 中可见的形态组织结构（细胞核形状、腺体结构、间质模式）转换为 IHC 突出显示的膜蛋白表达模式——这根本上是一个像素级的图像到图像翻译问题，而不是表示级别的对齐问题。相比之下，MoRA 的模态感知投影和 SimMLM 的门控在特征/令牌级别操作；当应用于图像+图像模态对时，它们必须跨越更大的语义差距，而不产生直接的像素级重建，使得更难验证适应的表示是否捕捉到区分 HER2 表达水平的精确

亚细胞染色模式。生成合成，尽管其质量较低，直接输出完整分辨率的 IHC 图像，供下游分类器利用——像素级的方法在结构上更符合该领域的信息结构，即使合成质量不完美。

最强的证据支持 MM'25 论文的生成方法是一个有意义的贡献，但可能不是缺失模态鲁棒性的最参数高效解决方案。对 CM-GAN 与 MoRA 风格适应与 SimMLM 风格门控在相同 HER2 分类任务上的比较评估将是决定性实验，文献表明多种方法可能共存——生成合成用于可解释性，参数高效适应用于效率。

4.2.2 主题：特征解耦——跨领域验证的架构

MM'25 论文中的病理信息解耦编码器——将共享特征（H&E 和 IHC 中的共同形态模式）与模态特定特征（H&E 中的结构模式；IHC 中的蛋白表达模式）分离——是最具架构意义的贡献之一，并且得到了 MODES（P008，npj Digital Medicine 2025）的独立验证。

为什么共享-特定解耦有帮助：互补信息假说。 解耦共享和特定特征的机制性理由在于 H&E 和 IHC 图像编码部分重叠但不相同的生物信息。共享特征捕捉两种模态的共同点：整体组织结构（腺体结构、间质分布、细胞核密度），反映了决定 H&E 外观和 IHC 染色模式的基础肿瘤形态。这些共享特征是跨模态转移的基础——它们是 H&E→IHC 合成可行的原因，因为存在可在两种染色中观察到的共同形态基质。特定特征捕捉每种模态独特揭示的信息：H&E 特定特征编码组织纹理、细胞核形态、建筑异常；IHC 特定特征编码膜蛋白表达强度、阳性染色细胞的空间分布以及病理学家用于 HER2 分级的膜完整性评分标准。通过分离这两条流，解耦编码器防止了干扰：IHC 特定的膜特征不会被共享路径中的 H&E 纹理模式稀释，而 H&E 特定的结构特征不会污染 IHC 特定的蛋白表达路径。MODES（P008）在完全不同的领域验证了这一原则——ECG 和心脏 MRI 共享心脏功能信息（心率变异性、射血分数），而每种编码独特信息（电活动与解剖结构）——确认互补信息假说是领域无关的，而非组织病理学特定的。

MM'25 论文通过 DCO（领域分类损失）和 DAO（分布对齐损失）从 Wang 等人 CVPR 2023 的实例化操作化了这一原则：DCO 通过训练领域鉴别器区分共享和特定特征分布，推动它们分开，而 DAO 将 H&E 和 IHC 的共享特征分布拉到对齐，以便它们是同一基础形态的可比较表示。然后，CBAM 启发的通道注意学习根据输入质量加权共享和特定的贡献——当真实的 IHC 模态可用时，IHC 特定路径提供高质量的膜信息，注意模块上调其权重；当仅有合成的 IHC 可用（H&E→IHC 重建）时，注意模块下调特定路径的权重，更依赖于共享的形态表示。这种自适应加权是注意模块贡献 +4.91% F1 改进的原因：它本质上是一个学习的门控机制，为每个输入选择最可靠的信息流，而不是使用固定的融合规则。

MODES 还提供了一个关键的架构见解：掩蔽组件（消除低信息表示维度）在不损失准确性的情况下保持准确性，同时减少表示维度。MM'25 论文的注意机制——在应用通道注意之前连接共享和特定特征——在概念上与这一发现一致：通道注意实际上是在学习掩蔽低价值特征通道。一个自然的后续步骤是将 MODES 风格的掩蔽明确地纳入 MM'25 框架，这可能改善表示的紧凑性，并可能提高泛化能力。

最强的证据是 MODES 的结果，即共享表示即使在源模态缺失时也能推断与模态相关的表型——这与 MM'25 论文通过共享表示路径仅从 H&E 预测 HER2 的能力完全相似，确认共享特征流是缺失模态鲁棒性的主要机制。

4.2.3 主题：跨模态重建质量及其与分类准确性的关系

文献中最反直觉的发现是跨模态重建质量（SSIM = 0.39–0.51）与下游分类准确率（94.25%）之间的解离。这一解离的重要性在于它对临床部署具有直接影响。

为什么 CNN 能容忍低 SSIM 合成：特征提取假说。 BCI 论文（P012）提供了关键证据，并进行了具体量化：病理学家在解释生成的 IHC 图像时仅获得 37.5% 和 40.0% 的准确率——确认生成的图像在像素级别上确实具有误导性，即使它们包含足够的判别信息供 CNN 分类器使用。生成图像的 SSIM（BCI 为 0.431，CM-GAN 为 0.51）远低于通常认为可接受的自然图像合成 SSIM > 0.90 范围，而病理学家的准确率确认这些低 SSIM 图像确实会让人类专家感到困惑，而不仅仅是美学上的不完美。

CNN 仍然保持高准确率的机制性解释根植于 CNN 特征提取的性质：下游 ResNet50 分类头在高层卷积特征图上操作，而不是像素级图像内容。CNN 学习层次表示，早期层捕捉纹理和边缘模式，而更深层捕捉抽象语义特征（细胞核形态、腺体结构、膜染色强度）。H&E→IHC 跨模态重建捕捉了足够的结构信息——膜区域的近似形状、阳性与阴性染色细胞的相对密度、整体组织结构——使得 CNN 特征提取器能够分类 HER2 表达水平，即使像素级细节（确切的膜棕色染色强度、精确的亚细胞定位）被错误呈现。这类似于 JPEG 压缩质量为 30 的图像仍然与原始图像几乎相同的分类准确率：CNN 对视觉感知层面的分布噪声具有鲁棒性，因为它们的决策边界是在平滑像素级伪影的特征空间中雕刻的。关键是，这种容忍是 CNN 架构的特征，而不是对领域转移的鲁棒性保证——问题在于这种容忍是否在伪影模式变化时保持（跨机构部署），这就是为什么 BCI 病理学家的证据确认了单一机构数据上的伪影问题，但未能验证跨机构鲁棒性。

为什么跨机构变异是关键威胁：伪影利用假说。 染色标准化基准（P010）通过具体的定量证据增加了另一个分析层次：即使是基于深度学习的染色标准化方法（CycleGAN、Pix2pix）也会在组织病理学图像上产生幻觉伪影——CycleGAN-ResNet 在脂肪组织中生成假细胞核；Pix2pix-DenseUNet 产生错误着色的平滑肌细胞核。传统方法产生不同的伪影：Macenko 在伊红中产生蓝色伪影；Vahadane 在许多情况下完全丢弃苏木精。P010 中的颜色转移数据量化了这一幅度：直方图匹配在皮肤上实现交集=0.891，但 CycleGAN 的 FID=61.67，表明结构失真显著。关键是，没有单一的标准化方法在所有组织类型和所有指标上占主导地位，这意味着 CM-GAN 的跨模态合成可能在 H&E 和 IHC 染色之间的关系不那么可预测的组织区域中同样产生幻觉。

这对临床有效性的影响机制有两个方面。首先，当 CM-GAN 在 BCI 单一机构上训练时，它不仅学习了 H&E 组织结构与 IHC 蛋白表达之间的真实形态关系，还学习了机构特有的特点：特定扫描仪的特征、特定染色协议的颜色特征、在一个机构内一致但在其他地方缺失的组织准备伪影。病理信息解耦编码器学习的共享表示捕捉了真实的跨模态形态以及这些机构特有

的模式。其次，当在新机构部署时，来自新机构的 H&E 图像具有不同的扫描仪特征和染色化学——共享编码器对这些不同输入的响应是未知的。如果新机构的组织伪影恰好与 BCI 训练伪影相似，跨机构准确率可能保持高（H1a）。如果它们有显著不同，共享表示的学习特征可能无法正确激活，分类头可能无法补偿（H1b）。P010 的证据表明，即使是同一组织块在 66 个实验室染色也会产生显著的视觉差异，表明 H1b 是针对单一机构数据训练的模型更可能的结果——"伪影利用"问题不是假设，而是基于跨机构染色变异的幅度的实证基础。

多目标染色标准化方法（P011）提供了潜在的缓解措施：使用多个参考图像进行标准化，而不是单一典型参考，这可以通过避免对单一染色颜色分布的过拟合来降低幻觉伪影的发生率。在 CM-GAN 之前将多目标标准化作为预处理步骤进行整合，可能是一个有价值的实验，专门用于解决伪影利用假说。

4.2.4 主题：跨机构泛化——关键有效性威胁

MM'25 论文在 BCI 数据集上的单一机构评估是最显著的有效性威胁，而染色标准化文献（P010）提供了最具体的证据，说明这为何重要。66 个实验室的多中心研究表明，即使是同一组织块在 66 个不同实验室染色也会产生显著的视觉差异：直方图匹配在参考染色上的交集得分范围从 0.891（皮肤）到 0.944（肾脏），而深度学习方法的 FID 得分范围则更高。关键是，没有单一的染色标准化方法在所有组织类型和所有指标上占主导地位——直方图匹配在皮肤和肾脏上表现最佳，而 Macenko 和 Reinhard 在结肠上表现相当。这意味着在一个机构的染色协议上训练的模型在部署到另一个机构时面临根本不同的输入分布，即使基础组织生物学是相同的。

ResViT-GANNet 论文（P005）通过具体数字强化了这一担忧：其双分支多模态架构（CNN + ViT）在 BreakHis 上实现 96.40% 的准确率，证明多模态组织病理学分类是广泛可行的——但 BreakHis 本身是一个多机构数据集，且不同放大倍数和肿瘤类型的准确率变异（8 种肿瘤类型 × 4 种放大倍数）表明对数据分布的敏感性。GAN 增强在少数肿瘤类别上带来的 3.3% 准确率提升表明数据多样性对困难病例的分类性能有重要影响。

HER2 MRI 多模态论文（P002）提供了一个有趣的反例：ResNet50 + 临床列线图在 MRI 特征与临床元数据的后期融合中实现 AUC=0.94，证明了从非组织病理学成像中进行 HER2 预测是可行的，并且 HER2 MRI 多模态论文 (P002) 提供了一个有趣的对比：ResNet50 + 临床列线图通过晚期融合 MRI 特征与临床元数据实现了 AUC=0.94，表明从非组织病理学成像中预测 HER2 是可行的，并且结合临床变量增加了区分价值。这暗示了 MM'25 框架的潜在扩展——将组织病理学特征与临床元数据（肿瘤分级、Ki-67 指数、激素受体状态）结合起来，可能在跨机构变异下提高鲁棒性。

MM'25 论文的直接含义是，94.25% 的 H&E 仅准确率应在不同染色协议的外部数据集上进行验证，然后才能进行临床部署。BCI 数据集的单一机构来源意味着报告的性能对于跨机构部署可能过于乐观。

4.2.5 主题：可解释性和不确定性 — 临床部署差距

文献中识别出两个关键的临床部署要求，但在 MM'25 论文中未得到解决：可解释性和不确定性量化。这些差距是相互关联的——没有经过校准的置信度估计，临床医生无法可靠地区分高置信度的正确预测和需要人工审核的不确定预测；而没有可解释的注意力图，即使是自信的预测也无法向病理学家或患者解释。

可解释性：注意力机制缺乏膜级定量验证。 HER2 可解释评估论文（P001）建立了 MAP（膜激活精度）和 EC（解释一致性）作为 HER2 IHC 评分可解释性的具体定量指标，评估是基于病理学家标注的膜掩模。混合解冻 ResNet50 的 MAP = 79%得分表明，模型的 79%的注意力与病理学家标注的膜区域重叠——这意味着大约五分之一的模型关注点落在病理学家认为与 HER2 评分在形态上无关的组织区域。MM'25 论文的 CBAM 启发的注意力模块在架构上被定位为生成类似的可解释性图，先连接共享和特定特征，然后进行通道注意力，这在概念上与 P001 的 Score-CAM 管道中的注意力机制是一致的。然而，MM'25 论文中没有报告 MAP/EC 风格的评估——注意力权重被可视化但没有与病理学家标注的真实值进行验证。这在临床上是重要的，因为 MAP = 79%（P001）是可以接受的但不是最佳的；注意力对齐的改进可能直接转化为更具临床可信度的解释。此外，P001 的 WSI 级结果显示 HER2 1+的精度明显低于 BCI 的补丁级结果（P=0.82, R=0.83，P=0.95, R=0.90, F1=0.85），这表明注意力质量在 WSI 尺度上下降——MM'25 论文的评估是在 BCI 的补丁级，因此其在 WSI 尺度上的注意力模块质量尚不清楚。

不确定性量化：没有校准，就没有临床信任。 SimMLM（P007）提供了多模态深度学习中经过校准的不确定性的最强定量证据：logit 级 DMoME 门控机制在 BraTS 2018 上实现了 ECE = 3.15%/3.75%/3.55%（ET/TC/WT），而基线 MoMKE 方法为 3.46%/4.10%/4.04%——校准误差减少了 0.31–0.35 个百分点。关键是，SimMLM 在三种 MRI 序列中实现了置信度可靠性（CR）减少 50.68%/13.18%/61.97%，这意味着模型的置信度估计在实际正确性方面具有显著的预测能力。这是将不确定性量化转化为临床效用的具体机制：不仅是更低的 ECE，还有更高的置信度可靠性。MM'25 论文没有提供置信度估计——这是临床部署的一个关键差距。P003 和 P004 明确指出不确定性量化的缺失是 HER2 AI 系统临床部署的三大障碍之一，其他障碍包括染色变异和有限的数据多样性。临床影响是严重的：在 HER2 1+病例中，观察者间一致性最低（根据 P003，2+病例为 60–75%，在 0/1+边界上显著更低），模型的预测最不确定，但治疗后果最为重要——T-DXd 的资格取决于正确的 HER2 低分类。没有经过校准的置信度估计，临床医生无法系统地识别哪些 HER2 1+预测需要病理学家的升级。MM'25 框架的具体校准目标是根据 P007 在 BraTS 的证据，ECE < 5%，对于 HER2 1+特别要求每类 ECE < 8%（因为 HER2 1+类别是最不确定且治疗风险最高的）。

可解释性和不确定性之间的相互联系。 P001 证明期望校准误差（ECE）是可解释性评估框架的一部分——评估框架包括 ECE 以评估模型的置信度可靠性，此外还有 MAP 和 EC。这意味着改善不确定性量化和改善可解释性注意力图并不是独立的目标；它们共享相同的校准质量：一个具有良好校准置信度（低 ECE）和高 MAP 的模型会产生既可解释又可信的预测。MM'25 论文的 CBAM 启发的注意力模块是这两个目标最自然的整合点——如果通道注意

力可以通过一个关注校准的损失进行训练（遵循 SimMLM 的 MoFe 原则，λ=0.1），并通过 MAP 与病理学家标注的膜区域进行评估，它可以同时解决在这一主题中识别出的可解释性和不确定性量化的差距。

基础模型背景。 CPath-Omni（P009）建议了一条更长期的路径：一个具有 150 亿参数的基础模型在 42 个计算病理基准中实现了 SOTA，接近人类病理学家的表现（专家 72.3% vs. CPath-Omni 71.8%在 PathMMU 上有 0.6%的优势）。然而，CPath-Omni 并未解决缺失模态的鲁棒性，表明这一特定能力在基础模型规模下仍然是一个开放的研究问题——验证了 MM'25 论文的具体贡献。CPath-Omni 中缺乏不确定性量化也表明，这在计算病理的前沿仍然是一个开放的挑战。

5. 当前项目的综合评估

来自基础来源和文献的证据支持以下 HER2 预测框架的综合评估：

强烈合理的方向：双模态灵活架构是由真实的临床需求（资源有限环境中的缺失模态）所驱动，并得到了来自三个独立研究方向的趋同证据的支持（根据 P012 的生成合成、根据 P008 的特征解耦、根据 P007 的动态门控）。特征解耦方法通过 MODES（P008，npj Digital Medicine 2025）在不同的临床领域（心血管成像与 ECG 和 cMRI）中得到了独立验证，显著增强了对架构设计的信心。CM-GAN 在 H&E 单模态基线上的 22.81%的改进确认了跨模态重建增加了实质性的区分价值，而不仅仅是表面的增强。受 CBAM 启发的注意模块的+4.91%的 F1 改进与 SimMLM 在 logit 级别模态加权上的发现一致（P007）。ResViT-GANNet（P005）在多机构 BreakHis 数据上实现的 96.40%的准确率确认了在真实世界多机构条件下可实现约 95%的准确率范围。

为什么 SSIM-准确率的分离在 BCI 上成立但可能不具普遍性。 CM-GAN 尽管 SSIM=0.51 却能实现 94.25%准确率的机制解释基于 CNN 特征提取假设：ResNet50 分类头在高层卷积特征上操作，这些特征能够容忍像素级合成伪影，因为与 HER2 分类相关的特征（整体膜区域密度、核-细胞质比率模式、腺体结构）即使在精确的亚细胞染色细节被错误呈现时也得以保留。这类似于为什么 CNN 对 JPEG 压缩在质量 30 时具有鲁棒性——特征空间的决策对人类观察者认为视觉上令人困惑的像素级噪声不敏感。BCI 病理学家对生成的 IHC 图像的 37.5-40%准确率（P012）确认了 SSIM=0.51 的图像确实对人类感知具有误导性，建立了在相同视觉输入上的人机性能差距。然而，这种容忍可能特定于 BCI 单机构设置，在该设置中，分类头已经学会依赖于特定于机构的伪影模式以及真实的形态特征。跨机构部署可能会暴露出鲁棒性是否普遍（H2：真实共享特征的鲁棒性）或崩溃（H1：特定于 BCI 训练分布的伪影利用）。

为什么跨机构变异是关键威胁。 跨机构部署为何是最严重的有效性威胁的机制有三个组成部分。首先，当 CM-GAN 在 BCI 的单一机构上训练时，它不仅学习了 H&E 组织结构与 IHC 蛋白表达之间的真实形态关系，还学习了机构的特性：特定扫描仪特征、特定染色化学、在一个机构内一致但在其他地方缺失的组织准备伪影。其次，病理信息解耦编码器的共享特征捕获了真实的跨模态形态以及这些特定于机构的模式——DCO 将共享特征和特定特征分开，但特定于机构的特征在训练机构的 H&E 和 IHC 图像中是共享的，因此它们渗入共享表示路径。第

三，P010 在相同组织块上来自 66 个实验室的证据表明，相同的生物组织在不同机构中产生截然不同的视觉输出：即使使用最佳方法，直方图匹配的颜色转移交集得分也仅为 0.891-0.944，这意味着 6-11%的染色颜色信息未对齐。当 CM-GAN 遇到来自新机构的 H&E 图像，其染色化学不同，共享编码器的响应是不可预测的——如果新机构的组织伪影类似于 BCI 训练伪影，准确率可能保持高（H1a）；如果它们有显著不同，共享表示可能会错误激活，跨模态合成质量将下降（H1b）。P010 的跨机构变异的幅度使得 H1b 在真实世界部署中更可能成为结果，使得伪影利用假设成为临床有效性的最关键威胁。

跨文献线索的综合合成：缺失模态鲁棒性的三种架构家族（根据 P012 的生成合成、根据 P006 的参数高效适应、根据 P007 的动态门控）并不是相互排斥的——它们解决了问题的不同方面。CM-GAN 优先考虑合成模态的视觉可解释性；MoRA 优先考虑计算效率；SimMLM 优先考虑校准。MM'25 论文的贡献在于这些交集的最强：它展示了跨模态特征的生成合成（CM-GAN）可以与共享-特定解耦（DCO/DAO）和基于注意的融合（CBAM）结合在一个端到端的管道中。这种组合是新颖的，并且在任何单一引用的论文中都没有找到。

MODES 和 SimMLM 如何共同约束设计空间：MODES（P008）确立了共享-特定解耦对于高质量多模态融合是必要的——基于连接的方法在准确率上留下了空间（在 MM'25 论文中，解耦相较于连接提高了+4.29%）。SimMLM（P007）确立了通过 logit 级别模态加权可以实现校准（ECE 降低 0.31–0.35 个百分点）。共同来看，这些发现表明 MM'25 框架的理想训练目标应当：（1）通过 DCO/DAO 强制实施共享-特定解耦（如当前所做），（2）增加一个校准感知的损失项（遵循 SimMLM 的 MoFe 原则，λ=0.1），并且（3）使用 ECE 指标评估不确定性。现有的 CBAM 注意模块在架构上非常适合校准扩展。

弱支持的假设：该框架的跨机构泛化完全未经验证。66 个实验室的染色标准化证据（P010）——表明即使是相同的组织块在 66 个不同实验室染色后产生的交集得分范围从 0.891 到 0.944，具体取决于组织类型和方法——表明在 BCI 数据集上报告的性能可能无法转移到其他机构。CM-GAN 的生成方法可能产生类似于 P010 中 CycleGAN 和 Pix2pix 所记录的幻觉伪影（脂肪组织中的假核、平滑肌核的错误着色）。HER2 1+的表现尚不清楚——94.25%的总体准确率可能掩盖了在 HER2 低边界的性能下降，而这是根据 P003 所述的临床上最重要且最困难的类别。

应谨慎对待的假设：尽管重建 SSIM 较低，但高分类准确率不应被解读为重建质量不重要的证据。分类头可能在利用特定于数据集的伪影（BCI 注册模式、机构染色特性）而非真实的 HER2 相关形态特征，在这种情况下，跨机构部署可能会看到比单一机构评估所暗示的更大的准确率下降。BCI 病理学家的证据（对生成图像的 37.5–40%准确率，P012）确认生成的图像在当前 SSIM 水平下并不具有人类可解释性——这限制了该框架在人工智能与人类混合工作流程中的临床实用性。CM-GAN 以 94.25%的准确率实现假 IHC 并不意味着假 IHC 图像在医学上有意义；这意味着分类器容忍它们的缺陷。

当前证据无法得出的结论：我们不能得出框架已经准备好进行临床部署的结论，除非在多机构数据上进行外部验证。我们不能得出 CM-GAN 优于 MoRA 风格或 SimMLM 风格缺失模态方法的结论，除非在相同的 HER2 分类任务上进行直接比较基准测试。我们不能得出框架在 HER2 0/1+边界上稳健处理的结论，除非进行每类准确率报告。我们不能得出 SSIM-准确率分离反映真实共享特征鲁棒性（H2）或数据集特定伪影利用（H1）的结论，除非有跨机构证据。

6. 未解决的问题和决策关键缺口

1. **跨机构泛化**：该框架仅在来自单一机构的 BCI 数据集上进行评估。66 个实验室的染色标准化研究（P010）表明，不同机构之间的染色变化是显著的，无法通过单一的标准化方法解决。颜色转移交集分数范围从 0.891（皮肤的直方图匹配）到 0.944（肾脏的直方图匹配），表明即使是最佳方法也会在不同机构之间留下约 5-10%的颜色信息未对齐。在临床部署考虑之前，必须在来自至少 2-3 个不同机构、使用不同扫描仪和染色协议的数据集上进行外部验证实验。

2. **HER2 1+ / HER2-low 边界性能**：最具治疗重要性和诊断难度的类别仅被隐式评估。综合 94.25%的准确率可能掩盖了 HER2-low 边界的性能显著下降，正如 P001 论文中发现的，BCI 10×上 HER2 1+的 F1 = 0.85（与 HER2 3+的 F1=0.99 相比），以及在 HER2-IHC-40x-WSI 上 Class 1+的 P=0.82，R=0.83。需要进行逐类混淆矩阵分析。P003 中 2+病例的观察者间一致性为 60-75%，强调了即使对于人类专家，这一边界确实很困难。

3. **不确定性量化**：该框架未提供置信度估计，这对于临床决策支持几乎是必需的要求。临床医生需要知道不仅是预测类别，还要知道预测正确的概率，特别是在边界案例中。SimMLM（P007）提供了一个具体的校准目标：在 BCI 数据集上 ECE < 5%。蒙特卡洛丢弃或温度缩放可以产生校准的概率；当病理学家审查不确定预测时，低于 80%置信度的修正率可以作为验证指标。

4. **可解释性评估**：受 CBAM 启发的注意力模块的输出未使用 MAP 和 EC 等指标与病理学家注释的膜区域进行评估。没有这一评估，尚不清楚注意机制是否产生临床有意义的解释。P001 提供了具体目标：MAP > 75%，EC > 80%。将 Score-CAM 应用于注意力模块输出，并将注意区域与病理学家注释的膜掩膜进行比较，将解决这一缺口。

5. **CM-GAN 与其他缺失模态方法的比较**：三种不同的架构家族已独立验证，但从未在同一 HER2 分类任务上进行基准测试。MoRA 风格或 SimMLM 风格的方法可能以更低的计算成本实现可比或更优的缺失模态鲁棒性（MoRA 每 P006 仅使用 1.6%的可训练参数）。在 BCI 数据集上的比较基准测试将确定生成合成的计算成本是否合理。

6. **重建质量与跨机构鲁棒性之间的关系**：尽管重建 SSIM 较低，MM'25 论文的高分类准确率可能特定于单一机构的 BCI 数据集。如果分类头利用数据集特定的伪影，跨机构部署可能会导致准确率显著下降。P010 提供了机制：CycleGAN 和 Pix2pix 产生幻觉伪影（脂肪组织中的假核、误色的平滑肌核），而 CM-GAN 可能在跨机构染色变化下产生类似的伪

影。重建 SSIM 与领域转移下分类准确率下降之间的相关性分析将确定合成质量是否预测跨机构鲁棒性。

7. **对其他免疫组化生物标志物的泛化能力**：该框架仅针对 HER2 进行了验证。HER2 MRI 论文（P002）表明，来自不同成像模态（MRI + 临床列线图）的多模态 HER2 预测实现了 AUC=0.94，表明多模态融合方法在成像模态之间具有泛化能力。然而，尚不清楚 CM-GAN 跨模态合成是否特别泛化到其他生物标志物预测任务（ER/PR，Ki-67），或者 HER2 特有的特征（强膜染色、明确的表达水平）是否使其特别适合这种方法。

8. **临床工作流程集成**：该框架假设可用 WSI 级别的输入，但未解决在全切片图像中选择感兴趣区域（ROI）、组织分割或处理超出组织区域的实际挑战。P001 提供的证据表明，冷冻骨干方法在 BCI 上从 95%的准确率严重下降到 73%（在 10×下），这表明骨干选择和微调策略对鲁棒性至关重要。使用自动组织分割和 ROI 检测的端到端评估将解决这一缺口。

7. 推荐的下一步措施

以下所有建议均按临床影响优先级排序，并基于文献中的具体证据。每项建议的结构包括：(1) 具体的行动；(2) 测量指标；(3) 目标值；以及 (4) 比较基线。

1. **进行跨机构外部验证**（最高优先级，针对每个 P010 的单机构有效性威胁）：

- **要做什么**：在 BCI 单机构数据集上训练 CM-GAN 框架；在来自不同扫描仪、染色协议和组织制备方法的 2-3 个外部 HER2 数据集上进行评估。 - **测量指标**：每个机构的准确性，从单机构到多机构平均的准确性下降。 - **目标值**：从 BCI 基线（94.25% H&E-only；95.09% dual-modality）下降的准确性小于 5 个百分点。 - **比较基线**：BCI 单机构 94.25%（H&E-only），95.09%（dual-modality），参考文献表 1。使用 P010 的 66 个实验室交集分数（0.891–0.944）作为预期染色变化幅度的参考。

2. **报告每类 HER2 性能与混淆矩阵**（针对每个 P001、P003 的 HER2-low 评估差距）：

- **要做什么**：在评估框架中添加分层的每类精确度、召回率、F1、AUC，以及完整的 4×4 混淆矩阵，使用类别平衡抽样以确保充分的 HER2 1+表示。 - **测量指标**：每个 HER2 类别（0, 1+, 2+, 3+）的 F1 分数。 - **目标值**：HER2 0 F1 ⩾ 0.95；HER2 1+ F1 ⩾ 0.85（与 P001 的 F1=0.85 在 BCI 10×匹配）；HER2 2+ F1 ⩾ 0.90；HER2 3+ F1 ⩾ 0.95。 - **比较基线**：P001 的每类 BCI 结果在 10×放大下 — 类别 0（P=0.97, R=0.97），类别 1+（P=0.95, R=0.90, F1=0.85），类别 2+（P=0.94, R=0.98），类别 3+（P=0.97, R=0.94）。

3. **将不确定性量化与校准概率输出结合**（针对每个 P003、P004、P007 的临床部署差距）：

- **要做什么**：实施蒙特卡罗丢弃法，进行 50 次以上的前向传递以生成基于集成的置信度估计；在训练后应用温度缩放以校准概率输出；为每个 HER2 类别生成可靠性图和置信度直方图。 - **测量指标**：在 BCI 验证集上的期望校准误差（ECE），按 HER2 类别细分。 - **目标值**：整体

ECE < 5%；HER2 1+类别的每类 ECE < 8%（最不确定的类别）。 - **比较基线**：SimMLM（P007）在 BraTS 上实现 ECE = 3.15%/3.75%/3.55%，置信度可靠性降低 50.68%/13.18%/61.97%。

4. **在 BCI 上实施并基准测试 MoRA 风格和 SimMLM 风格的缺失模态替代方案**（针对每个 P006、P007 的比较架构问题）：

- **要做什么**：在 BCI HER2 分类任务上实施 MoRA 风格的模态感知 LoRA 适配（秩 r=4，每个 P006 的可训练参数 1.6%）和 SimMLM 风格的 DMoME 门控（每个 P007 的 MoFe 损失系数 λ=0.1）；在相同的 BCI 训练/验证/测试分割上训练这三种方法；评估 H&E-only、IHC-only 和双模态的准确性。 - **测量指标**：H&E-only 准确性、IHC-only 准确性、双模态准确性、每类 F1、GPU 内存使用（GB）、每 1,000 步的训练时间（小时）。 - **目标值**：如果 MoRA 或 SimMLM 的准确性在 CM-GAN 的准确性之内 2 个百分点，则优先选择参数高效的方法以便于资源有限的部署。 - **比较基线**：CM-GAN 的当前结果 — 94.25% H&E-only，90.28% IHC-only，95.09% dual-modality，参考文献表 1。

5. **使用 MAP/EC 指标与病理学家验证评估 CBAM 注意力可解释性**（针对每个 P001 的可解释性差距）：

- **要做什么**：将 Score-CAM 应用于 CBAM 灵感的注意力模块输出；计算模型注意区域与病理学家标注的膜区域之间的重叠率 MAP；至少有 2 名获得认证的病理学家在 200 张 BCI 图像上标注膜区域。 - **测量指标**：MAP（模型注意与病理学家标注的膜区域重叠的百分比）；EC（在≥3 个相似输入之间注意区域的稳定性）。 - **目标值**：MAP > 75%（与 P001 的 MAP=79%匹配）；EC > 80%。 - **比较基线**：P001 的 ResNet50 混合解冻在 BCI 上实现 MAP=79%在 10×。

6. **在 CM-GAN 跨模态合成之前应用多目标染色归一化作为预处理**（针对每个 P010、P011 的幻觉伪影问题）：

- **要做什么**：在 CM-GAN 的 H&E→IHC 合成之前，将多目标染色归一化（P011）作为输入预处理步骤；测试 Macenko 和直方图匹配作为替代预处理策略；使用 66 个实验室的染色归一化基准数据集（P010）来评估重建伪影率。 - **测量指标**：重建伪影率（每位病理学家评估中可见幻觉伪影的图像百分比）；外部数据上的重建 SSIM 和 PSNR。 - **目标值**：在外部数据集上重建伪影率降低> 20%；在外部数据上保持或提高重建 SSIM ⩾ 0.45。 - **比较基线**：P010 记录 CycleGAN-ResNet 在脂肪组织中生成假核，Pix2pix-DenseUNet 在跨机构变化下产生错误着色的平滑肌核。

7. **使用混合解冻策略调查 HER2 0/1+二元边界性能**（针对每个 P001、P003 的 HER2-low 边界差距）：

- **要做什么**：进行针对 HER2 0 与 1+分类评估的重点二元分类（最具治疗重要性的边界 — T-DXd 资格每个 P003）；将 P001 的混合解冻训练策略应用于 MM'25 论文的 IHC 单模态分支；使用类别平衡抽样。 - **测量指标**：二元 HER2 0/1+ AUC-ROC；二元准确性；0/1+边界的每类精确度和召回率。 - **目标值**：在 BCI 上 HER2 0/1+二元 AUC ⩾ 0.92；IHC 单模态分支准确性 ≥ 90%。 - **比较基线**：P001 的 BCI 10× HER2 1+结果 — F1=0.85，P=0.95，R=0.90。IHC 单模态基线为 77.38%，参考文献表 1。

8. **使用混合解冻策略提高 IHC 单模态分支准确性**（针对每个 P001、P003/P004 与文献上限的 20 个百分点差距）：

- **要做什么**：将 P001 的混合解冻 ResNet50 微调策略应用于 MM'25 论文的 IHC 单模态分支（冻结块 1-2 以获取通用组织学特征，微调块 3-4 和全连接层）；实施：Adam 优化器（lr=0.001），余弦退火调度器，提前停止（耐心=5），L2 正则化（1×10^{-4}），数据增强（水平翻转，±20°旋转，颜色抖动），批量大小 32，最多 100 个周期。 - **测量指标**：BCI 验证集上的 IHC 单模态准确性；每个 HER2 类别的 F1；计算成本（每次训练运行的 GPU 小时数）。 - **目标值**：IHC 单模态准确性 ≥ 90%（与当前 77.38%相比，参考文献表 1），代表≥ 12.6 个百分点的改善；计算成本减少≥ 70%与完全微调相比（P001 记录 72.7%的减少）。 - **比较基线**：IHC 单模态基线为 77.38%（仅真实 IHC），参考文献表 1。文献上限为 97.9%用于 IHC 的幻灯片级 HER2 分类（每个 P003/P004）。P001 的 ResNet50 混合解冻在 BCI 上实现 95%的准确性在 10×。

8. 关键风险、警告和证据边界

1. **单机构评估风险**（引用来源：P010）：所有实验使用来自一个机构的 BCI 数据集。66 个实验室的染色标准化研究（P010）提供了具体证据，表明跨机构染色变异是显著的——直方图匹配在 66 个实验室之间的交集达到 0.891–0.944，这意味着即使使用最佳标准化方法，6–11%的颜色信息仍未对齐。报告的 94.25%的准确率可能无法推广到其他机构。任何临床部署决策必须等待外部验证结果。

2. **SSIM-准确性解离机制不确定性**（引用来源：P012）：CM-GAN 的跨模态 SSIM 值（0.39–0.51）显著低于 SSIM > 0.90。BCI 病理学家对生成的 IHC 图像的解读准确率为 37.5–40%（P012），证实了合成伪影在视觉上对人类专家具有误导性。低 SSIM 与高分类准确性之间的解离可能反映了伪影利用（H1）——在这种情况下，跨机构部署可能会因伪影模式的不同而导致准确性显著下降——或真正的共享特征鲁棒性（H2）。跨机构评估是唯一的决定性区分实验。

3. **缺失的不确定性量化风险**（引用来源：P003, P004）：该框架未产生置信度估计。在临床部署中，这意味着无法识别和升级对边界 HER2 2+或 HER2 1+病例的错误预测以供人工审查。P003 将缺乏不确定性量化识别为 HER2 AI 系统临床部署的三大主要障碍之一。P007 证明，在多模态设置中，通过专家的 logit 级混合可以实现校准（ECE < 5%）。

4. **HER2 低边界性能风险**（引用来源：P001, P003）：最具治疗重要性的类别（HER2 1+, 根据 P003 符合 T-DXd 标准）未单独评估。P001 在 BCI 上对 HER2 1+的 F1 值为 0.85，显著低于总体准确率。P003 报告对模棱两可的 2+病例的观察者间一致性为 60-75%，表明即使是人类专家也发现这一边界难以判断。MM'25 论文的总体准确率可能因在更容易的 HER2 0 和 HER2 3+类别上表现强劲而被夸大。

5. **来自非组织病理文献的证据转移风险**（引用来源：P006, P007, P008）：几篇关键支持论文（MoRA, SimMLM, MODES）是在非组织病理领域（CXR, ODIR 根据 P006; BraTS 根据 P007; ECG/cMRI 根据 P008）进行评估的。它们的架构原则直接转移到组织病理 H&E+IHC 图像+图像多模态学习的可行性尚未得到证明。MODES（P008）验证了跨领域的共享特定解耦原则，但具体实现（DCO/DAO 损失，CBAM 注意力）可能需要针对组织病理进行领域特定的调整。

6. **类别不平衡风险**（引用来源：P004, P005）：论文提到类别特定的训练权重，但未报告 BCI 数据集中 HER2 表达水平的分布。P004 将有限的训练数据多样性识别为 HER2 AI 临床部署的三大主要障碍之一。P005 的 ResViT-GANNet 在多机构的 BreakHis 上实现了 96.40%的准确率，但在少数肿瘤类别上通过 GAN 增强特别显示出 3.3%的准确率提升，表明类别不平衡对性能有实质性影响。

7. **生成模型评估限制**（引用来源：P010, P012）：CM-GAN 的跨模态重建仅在 BCI 单机构数据集上进行评估。P010 记录了基于深度学习的染色标准化方法（CycleGAN, Pix2pix）在 H&E 图像上产生幻觉伪影。CM-GAN 可能在跨机构染色变异下产生类似的伪影，但这一关系尚未被分析。BCI 生成的 IHC 的病理学家准确率为 37.5–40%（P012）证实了单机构数据上的伪影问题。

8. **未与最先进的 HER2 AI 系统进行比较**（引用来源：P001, P003, P004）：论文未与已建立的 HER2 AI 评分工具进行基准测试。P001 的混合解冻 ResNet50 在 BCI 上实现了 95%的准确率（与 MM'25 的 95.09%双模态相比），而 P003/P004 报告 IHC 的幻灯片级 HER2 分类准确率为 97.9%。单模态 IHC 基线的 77.38%显著低于文献中的最先进水平，表明 IHC 分支需要架构改进。

本报告基于对 ACM MM'25 论文（arXiv:2506.10006v2）及 12 篇支持文献的实证分析生成。