Research Report

Ground ID: pdf-2506.10006v2*20260416025902 **Source**: ACM MM'25 Paper — Dual-Modality Flexible HER2 Expression Prediction Framework **Literature Result**: `data/lit*results/pdf-2506.10006v2_20260416025902/lit.md` **Report Date**: 2026-04-16

1. Executive Overview

HER2 (Human Epidermal Growth Factor Receptor 2) expression prediction in breast cancer histopathology represents a high-stakes clinical task with significant implications for treatment decisions. The ACM MM'25 paper under review introduces a deep learning framework that achieves 95.09% accuracy with dual-modality histopathology inputs (H&E and IHC stained images) and maintains 94.25% accuracy when only H&E images are available, by leveraging a Cross-Modal Generative Adversarial Network (CM-GAN) to reconstruct the missing IHC modality from H&E. This dual-modality flexibility addresses a critical practical constraint: in many clinical settings, especially resource-limited environments, paired H&E and IHC images for the same patient tissue block are frequently unavailable due to cost, tissue availability, or laboratory workflow limitations.

The literature analysis reveals that this work sits at the intersection of three active research threads in computational pathology: missing-modality robustness in multimodal deep learning, cross-modal image synthesis for histopathology, and interpretable HER2 scoring automation. The most significant finding from the grounded analysis is that the framework's cross-modal reconstruction quality (SSIM = 0.39–0.51) is relatively low in absolute terms — yet the downstream classification accuracy remains high — suggesting that the classification head is robust to synthesis artifacts in ways that human pathologists are not (BCI pathologists achieved only 37.5–40% accuracy interpreting generated IHC images per P012). The second most significant finding is that the pathological information decoupling encoder's shared-specific feature decomposition approach is independently corroborated by MODES (P008, npj Digital Medicine 2025), which validates the decoupling principle across a completely different clinical domain (cardiovascular imaging with ECG and cardiac MRI), substantially strengthening the architectural design rationale.

The most critical unresolved issue is the single-institution evaluation. The 66-laboratory multicenter stain normalization benchmarking study (P010, Scientific Reports 2026) provides concrete evidence that even the same tissue block stained across different labs produces dramatic visual variation, making cross-institution generalization a non-trivial concern for deployment. Additionally, the framework provides no uncertainty quantification, which is a near-essential requirement for clinical decision support systems. The HER2-low (IHC 1+) boundary — the most therapeutically important and diagnostically most difficult

category — is evaluated only implicitly through aggregate accuracy, with no per-class breakdown reported.

2. Problem Setting and Source Context

2.1 Clinical Background: HER2 Expression in Breast Cancer

HER2 is a protein receptor on breast cancer cells that, when overexpressed (HER2 positive, IHC 3+ or IHC 2+ with positive ISH), indicates aggressive tumor behavior and eligibility for targeted therapies such as trastuzumab (Herceptin) and the antibody-drug conjugate T-DXd. More recently, HER2-low expression (IHC 1+ or IHC 2+ with negative ISH) has acquired therapeutic significance, as these tumors respond to T-DXd even though they were previously considered HER2-negative. This clinical development has made accurate HER2 scoring — particularly the HER2-low boundary — more important than ever. The grading scale ranges from HER2 0 (no expression), HER2 1+ (weak expression), HER2 2+ (moderate expression, requiring ISH confirmation), to HER2 3+ (strong overexpression).

The standard diagnostic workflow involves H&E (Hematoxylin and Eosin) staining for general morphology assessment, followed by IHC (Immunohistochemistry) staining that specifically highlights HER2 protein expression on cell membranes. In routine clinical practice, H&E and IHC images are rarely paired for the same tissue block, and their availability depends on separate tissue sections, staining runs, and laboratory schedules.

2.2 The Missing Modality Problem in Computational Pathology

Existing multimodal AI models in computational pathology typically assume complete modality availability — that is, paired inputs from multiple imaging modalities are available at inference time. This assumption fails in real clinical workflows for three reasons: (1) tissue scarcity — only one tissue section may be available for staining; (2) cost constraints — IHC staining is significantly more expensive than H&E; and (3) workflow asynchrony — H&E and IHC are often processed at different times and may come from different tissue blocks or biopsy sites.

The ACM MM'25 paper addresses this problem through a dual-branch architecture that dynamically selects either a cross-modal reconstruction path (for single-modality input) or an end-to-end fusion pipeline (for dual-modality input), based on a lightweight modality classifier. This design enables the framework to maintain high accuracy regardless of which modality is available, without requiring the expensive paired dataset for training.

2.3 Source Paper Scope and Contributions

The source paper makes four core contributions: (1) a Cross-Modal Generative Adversarial Network (CM-GAN) based on Pyramid Pix2Pix for bidirectional H&E↔IHC reconstruction in feature space; (2) a pathological information decoupling encoder using Domain Classification Loss (DCO) and Distribution

Alignment Loss (DAO) to separate shared and modality-specific features; (3) a CBAM-inspired modality-sensitive feature attention module for adaptive modality weighting; and (4) a comprehensive evaluation on the BCI (Breast Cancer Immunohistochemistry) dataset with 4,870 registered H&E-IHC image pairs.

3. Grounded Findings from the Source Material

3.1 Dual-Modality Flexible Architecture

The framework employs four core modules in sequence. First, a **branch selector** (99.95% classification accuracy) determines whether the input is single-modality or dual-modality and routes accordingly. Second, for single-modality inputs, the **CM-GAN** reconstructs the missing modality — this module uses a bidirectional GAN that can synthesize IHC features from H&E inputs and vice versa, built on the BCI Pyramid Pix2Pix architecture with modifications. Third, for both single- and dual-modality inputs, the **pathological information decoupling encoder** (pretrained ResNet50 backbone) separates shared features (morphological patterns common to both H&E and IHC) from modality-specific features (H&E captures tissue structure; IHC captures protein expression patterns), using DCO to push shared and specific features apart and DAO to align feature distributions across domains. Fourth, the **CBAM-inspired modality-sensitive attention module** concatenates shared and specific feature vectors before applying channel-wise attention, adaptively reweighting modality contributions based on input quality.

3.2 Performance Results

The key performance results reported in the paper are: dual-modality (Real H&E + Real IHC) achieves 95.09% accuracy (F1 = 0.9532); cross-modal (Real H&E + Fake IHC) achieves 94.25% accuracy (F1 = 0.9609), representing a +22.81% improvement over the H&E unimodal baseline of 71.44%; cross-modal (Real IHC + Fake H&E) achieves 90.28% accuracy (F1 = 0.9038), a +12.90% improvement over the IHC unimodal baseline of 77.38%; and the attention module contributes an additional +4.91% F1 improvement over the no-attention baseline.

3.3 Reconstruction Quality

Cross-modal reconstruction quality metrics: H&E-to-IHC reconstruction achieves PSNR = 18.48 dB, SSIM = 0.51; IHC-to-H&E reconstruction achieves PSNR = 17.24 dB, SSIM = 0.39. These SSIM values are relatively low in natural image synthesis literature — for context, the BCI Pyramid Pix2Pix baseline (P012) achieves SSIM=0.431 on the same task, against which CM-GAN's 0.51 represents approximately an 18% relative SSIM improvement, yet both remain well below the SSIM > 0.90 range typically considered high-quality for natural image generation.

3.4 Identified Constraints

The paper's own identified limitations include: single-institution evaluation (BCI dataset, one institution), no uncertainty quantification in model outputs, no per-class accuracy reporting (particularly absent for HER2 1+), no comparison with

state-of-the-art HER2 AI systems, class imbalance not quantified, and no ablation isolating DCO vs. DAO contributions.

3.5 Dataset and Architecture Details

The BCI dataset contains 4,870 registered H&E-IHC image pairs from a single institution, covering all four HER2 expression levels. The ResNet50 backbone is pretrained on ImageNet and fine-tuned with the hybrid unfreezing strategy (Blocks 1–2 frozen for generic features, Blocks 3–4 + fc fine-tuned). The CM-GAN uses multi-scale residual blocks with skip connections across four Gaussian pyramid levels, trained with combined GAN loss and L1 pixel loss across scales.

4. Literature-Based Deep Analysis

4.1 Preserved Detailed Paper Analyses

Paper 1: Interpretable and Robust Deep Learning for Automated HER2 Assessment (P001, Research Square, 2026)

Problem and Task Setting: The paper addresses three interconnected challenges in automated HER2 IHC scoring: limited interpretability of deep learning models (black-box problem), poor robustness across magnifications (10× vs 40×), and insufficient alignment with pathological reasoning. The task is 4-class HER2 expression classification (0, 1+, 2+, 3+) evaluated on three public datasets: BCI (10×), HER2-IHC-40x-Patch, and HER2-IHC-40x-WSI.

Methodology: The framework employs ResNet50 with three training configurations compared: Frozen Backbone (only fc layer trained), Full Fine-Tuning (all layers trained), and Hybrid Unfreezing (Blocks 1–2 frozen for generic features, Blocks 3–4 + fc fine-tuned). The hybrid approach is motivated by the observation that early ResNet layers capture generic histological structures (cell membranes, nuclei, tissue patterns) while later layers encode magnification-specific and task-specific features. For interpretability, Score-CAM generates attention maps, and two quantitative metrics are introduced: MAP (Membrane Activation Precision), measuring overlap between model attention and pathologist-annotated membrane regions, and EC (Explanation Consistency), measuring stability of explanations across similar inputs. The evaluation framework also includes Expected Calibration Error (ECE) to assess model confidence reliability. Training uses Adam optimizer (lr=0.001), cosine annealing scheduler, early stopping (patience=5), L2 regularization (1×10^{-4}), and data augmentation (horizontal flip, ±20° rotation, color jitter). Batch size is 32, 100 epochs max.

Main Evidence: The Hybrid Unfreezing configuration achieves 95% accuracy on BCI (vs. 73% frozen backbone, 91% full fine-tuning) and 96% on HER2-IHC-40x-Patch (vs. 91% frozen backbone, 95% full fine-tuning). The improvement from hybrid unfreezing is most pronounced on BCI at 10×, where full fine-tuning achieves only 91% vs. 95% for hybrid. Computational expense is reduced by 72.7% compared to full fine-tuning. Per-class BCI results: Class 0 (P=0.97,

R=0.97), Class 1+ (P=0.95, R=0.90), Class 2+ (P=0.94, R=0.98), Class 3+ (P=0.97, R=0.94). On HER2-IHC-40x-WSI: Class 1+ (P=0.82, R=0.83) — notably lower than BCI, suggesting sensitivity to dataset distribution. MAP score of 79% indicates that 79% of the model's attention overlaps with pathologist-annotated membrane regions. AUC > 0.99 for HER2 3+ discrimination.

Relevance: The hybrid unfreezing approach is architecturally complementary to the MM'25 paper's pretrained ResNet50 backbone — the MM'25 paper could benefit from the layer-freezing strategy to reduce computational cost while maintaining performance. The MAP/EC metrics provide a concrete evaluation framework for assessing whether the MM'25 paper's attention module produces clinically meaningful explanations. The BCI dataset results at 10× directly demonstrate the magnification robustness challenge: frozen backbone drops to 73% on BCI, suggesting that the MM'25 paper's single-magnification evaluation on BCI patches may not fully characterize cross-magnification generalization.

Limits: Focuses exclusively on IHC images (not H&E+IHC multimodal) and does not address missing modality scenarios. Class 1+ F1=0.85 at BCI 10× reflects the known difficulty of the HER2-low boundary. Formal pathologist validation of MAP/EC metrics was not conducted — only automated overlap with pathologist annotations. Single-institution dataset origin (BCI), same as the MM'25 paper.

Paper 2: HER2 Expression Prediction with Multimodal MRI (P002, Frontiers in Oncology, 2025)

Problem and Task Setting: Addresses HER2 prediction from MRI imaging (T1, T2, contrast-enhanced sequences) combined with clinical nomogram features, for patients where IHC results are unavailable or inconclusive. This is a fundamentally different imaging modality from histopathology.

Methodology: Deep learning (ResNet50, VGG16, EfficientNet-B0, ViT-Small) combined with clinical nomogram features using late fusion. ICC filtering and LASSO regression for feature selection. AUC-ROC as primary metric. Three-class classification (HER2 positive, HER2 negative, equivocal). The late fusion approach involves independently training each imaging modality backbone, extracting features from the final fully connected layer, concatenating them with clinical nomogram features (tumor size, grade, Ki-67 index, etc.), and passing the combined feature vector through a classification head.

Main Evidence: AUC = 0.94 using ResNet50 + clinical data integration. The late fusion approach (ResNet50 + clinical nomogram) achieves the highest AUC, outperforming individual modality backbones and imaging-only fusion. Demonstrates that HER2 prediction from non-IHC imaging modalities is feasible and that combining structural imaging features with clinical metadata provides complementary diagnostic information.

Relevance: Corroborates the MM'25 paper's premise that missing-modality HER2 prediction is clinically valuable. The late fusion architecture (P002)

represents an alternative to the MM'25 paper's shared-specific decoupling approach: P002 concatenates features from different modalities at the decision level, whereas the MM'25 paper decomposes features into shared and specific streams before fusion. The P002 finding that clinical nomogram features (tumor grade, Ki-67) add discriminative value beyond imaging alone suggests a potential extension for the MM'25 framework — incorporating clinical metadata alongside histopathology features could improve HER2 classification robustness. The different imaging modalities (MRI vs histopathology) provide complementary biological information, though the MM'25 paper's specific contribution remains the H&E+IHC histopathology multimodal fusion approach.

Limits: Different imaging modality from the MM'25 paper; MRI-based HER2 prediction is not yet clinically validated for this specific task. The late fusion approach is architecturally different from the MM'25 paper's feature decoupling encoder and has not been validated on the same HER2 classification task.

Paper 3: HER2-low Breast Cancer — Challenges in Identification, Detection, and Treatment (P003, PMC, 2024)

Problem and Task Setting: HER2-low (IHC 1+ or 2+ with negative ISH) represents ~40-50% of breast cancers and is therapeutically targetable with T-DXd, making accurate identification critical. The challenge is that HER2-low exists on a continuum with HER2 0, making the 0/1+ boundary the most diagnostically difficult.

Methodology: Systematic review of challenges across pre-analytical (tissue fixation, processing), analytical (antibody clone, scoring guidelines), and post-analytical (interpreter variability) factors. Intratumoral heterogeneity is identified as a major contributor to HER2 status variability. The review synthesizes evidence from clinical studies, pathology guidelines, and emerging computational pathology literature.

Main Evidence: The greatest impact of methodological variability is at the IHC 0-1+ interface. Dynamic nature of HER2-low status: HER2 expression can change following systemic therapy, requiring biomarker reassessment. Inter-observer concordance at the 0/1+ boundary is reported at 60-75% for equivocal (2+) cases, and substantially lower for borderline cases. Computational pathology is identified as a promising direction but not yet clinically validated. The review notes that inter-observer variability in HER2 scoring is a well-documented problem, with the highest disagreement occurring at the HER2-low boundary.

Relevance: Provides clinical context for the MM'25 paper's motivation and quantifies the magnitude of the diagnostic challenge. The 0/1+ boundary is the most variable region — with inter-observer concordance of 60-75% for 2+ cases per the systematic review — directly relevant to the MM'25 paper's evaluation. The MM'25 paper does not report per-class accuracy, leaving open whether the framework's 94.25% aggregate accuracy hides degraded performance specifically for HER2 1+ cases. The clinical imperative for accurate HER2-low

detection (therapeutic eligibility for T-DXd) strengthens the urgency of addressing this gap. The finding that HER2 expression can change following systemic therapy also raises a longitudinal validation consideration that the MM'25 paper's cross-sectional evaluation does not address.

Limits: Review article; does not propose computational solutions. The specific performance numbers cited (60-75% concordance) are for equivocal 2+ cases rather than the 0/1+ boundary specifically.

Paper 4: Computational Pathology in HER2-low Identification (P004, PMC, 2024)

Problem and Task Setting: Identifies opportunities and challenges for computational pathology in HER2-low identification, focusing on the gap between research prototypes and clinically deployable systems. Analyzes published AI approaches to HER2 scoring to assess the field's readiness for clinical deployment.

Methodology: Analysis of published AI approaches to HER2 scoring, identifying three major gaps: (1) lack of standardized evaluation protocols for HER2-low; (2) limited training data diversity; (3) absence of uncertainty quantification in model outputs. The paper surveys studies including deep learning classifiers, attention-based models, and foundation model approaches, assessing their performance on HER2-low cases.

Main Evidence: No single published computational pathology approach has been validated for HER2-low classification in routine clinical practice. Stain variation, tissue heterogeneity, and reader variability are identified as the three largest barriers to clinical deployment. The literature review finds that most published HER2 AI studies report aggregate accuracy but fail to provide per-class performance breakdown, particularly for HER2 1+. The review also notes that existing HER2 AI systems have not been prospectively validated in clinical settings.

Relevance: The three identified gaps map directly to limitations of the MM'25 paper: single-institution BCI dataset (data diversity gap), no per-class performance reporting (no HER2-low-specific evaluation), and no uncertainty quantification (clinical deployment gap). The finding that HER2-low identification is an unresolved challenge across the entire field — not just for the MM'25 paper — contextualizes the paper's contribution: it addresses an important problem that the broader community has not yet solved. The paper's identification of stain variation as the largest deployment barrier directly corroborates the single-institution validity concern raised by the MM'25 paper's evaluation.

Limits: No new experimental results; relies on literature synthesis. The specific barrier quantification (stain variation as "largest") is qualitative rather than quantitative.

Paper 5: ResViT-GANNet — Multimodal Attention and GAN-Based Augmentation for Breast Cancer Histopathology (P005, BMC Medical Imaging, 2025)

Problem and Task Setting: Classification of breast cancer histopathology images using a multimodal architecture combining CNNs with Vision Transformers (ViT), Token-Aligned Multimodal Attention (TAMA), and StyleGAN2-ADA for synthetic data augmentation. Validated on the multi-institutional BreakHis dataset covering 8 tumor types and multiple magnification factors.

Methodology: Dual-branch architecture: CNN branch (ResNet50/EfficientNet backbone) for H&E images + ViT branch for additional imaging data. TAMA module aligns and fuses heterogeneous features from both branches using cross-attention between CNN and ViT token sequences. StyleGAN2-ADA generates synthetic histopathology images for data augmentation during training. The framework processes images at 8 magnifications (40×, 100×, 200×, 400×) and 8 tumor types (breast cancer subtypes).

Main Evidence: 96.40% accuracy on breast cancer histopathology classification with a 3.3% improvement from synthetic data incorporation. TAMA module outperforms late fusion in multimodal settings, achieving 2.1% higher accuracy on BreakHis. The ablation study shows that GAN augmentation contributes 3.3% accuracy improvement specifically on minority tumor classes. Importantly, ResViT-GANNet is validated on BreakHis, which is a multi-institutional dataset with inherent staining variation across contributing institutions — making the 96.40% accuracy a more robust performance estimate than single-institution results.

Relevance: Shares the dual-branch multimodal architecture concept with the MM'25 paper and validates this architectural family on multi-institutional data. The TAMA module is architecturally similar to the MM'25 paper's CBAM-inspired modality-sensitive attention — both dynamically weight multimodal features using cross-attention mechanisms. The 96.40% accuracy on multi-institutional BreakHis is particularly significant: it demonstrates that the ~96% accuracy range is achievable under real-world multi-institutional conditions, suggesting that the MM'25 paper's 95.09% dual-modality accuracy on single-institution BCI is in a plausible performance range for deployment. The GAN augmentation finding (3.3% improvement) is conceptually related to the MM'25 paper's CM-GAN cross-modal synthesis, suggesting that generative models have measurable value in histopathology classification. However, ResViT-GANNet does not address missing modality scenarios — it assumes both modalities are available — so it cannot directly replace CM-GAN's missing-modality robustness contribution.

Limits: Does not address missing modality scenarios; does not evaluate on the BCI dataset or HER2-specific tasks; synthetic augmentation and cross-modal reconstruction are conceptually related but different tasks. The BreakHis dataset uses different tumor types rather than HER2 expression levels, limiting direct comparability to the MM'25 paper's task.

Paper 6: MoRA — LoRA Guided Multi-Modal Disease Diagnosis with Missing Modality (P006, arXiv, 2024)

Problem and Task Setting: Multi-modal pre-trained models degrade sharply when modalities are missing, and full fine-tuning is computationally expensive. MoRA addresses both challenges by using modality-aware low-rank adaptation on the first transformer block of a pre-trained multimodal model.

Methodology: MoRA projects each input to a low intrinsic dimension with modality-aware up-projections. Instead of shared LoRA projections for all modalities, MoRA uses different up-projection matrices for different modality availability scenarios (complete, modality-1-missing, modality-2-missing). When modality m1 is missing, the corresponding projection for that scenario is selected. The key innovation is that only 1.6% of parameters are trainable vs. full fine-tuning. Rank r=4 was found to be optimal on ODIR. The approach integrates into the first transformer block because the first layer can directly obtain information about the input token, helping confirm the status of missing modularity. Datasets: CXR (3,030 train, 385 val, 379 test, 20 diseases) and ODIR (2,781 train, 382 val, 337 test, 7 diseases).

Main Evidence: On CXR, at 30% image + 100% text missing rate, MoRA achieves F1-Macro=37.22 vs. MAPs=33.49 vs. MSPs=35.13 vs. ViLT=25.54. On ODIR, at 30% image + 100% text, MoRA achieves F1-Macro=92.56 vs. MAPs=90.66 vs. MSPs=46.38 vs. ViLT=81.34. When text is the missing modality on ODIR (100% image, 30% text), MoRA achieves F1-Macro=76.89 vs. MAPs=78.71 — in this case MAPs slightly outperforms MoRA. GPU memory: MoRA requires 12.2 GB vs. MAPs 14.4 GB on CXR; training time 1.58h vs. MAPs 1.71h per 1,000 steps. Rank sensitivity: r=4 is optimal (F1=80.73); r=384 (full rank) gives 70.23 F1-Macro, worse than without MoRA, confirming that low-rank is essential.

Relevance: Provides the strongest alternative approach to missing modality robustness without generative image synthesis. The modality-aware up-projection mechanism is conceptually related to the MM'25 paper's branch-selector + CM-GAN pipeline: both select different processing paths based on modality availability. However, MoRA requires pre-trained multimodal transformers while the MM'25 paper trains from scratch. Comparative benchmarking of CM-GAN vs. MoRA on the same HER2 classification task would clarify whether CM-GAN's generative approach is superior. MoRA's parameter efficiency (1.6%) contrasts with CM-GAN's full generation pipeline — if MoRA achieves comparable accuracy on HER2 classification, it would represent a substantially more efficient solution.

Limits: Requires pre-trained multimodal models; not directly applicable to the MM'25 paper's training-from-scratch setting; no histopathology-specific evaluation. The approach is evaluated on image+text modalities (CXR, ODIR), not image+image (H&E+IHC), so direct transferability to the MM'25 paper's domain is unproven.

Paper 7: SimMLM — Multi-modal Learning with Missing Modality (P007, ICCV 2025)

Problem and Task Setting: Existing missing modality methods rely on complex architectures or data imputation. SimMLM provides a generic framework with a Dynamic Mixture of Modality Experts (DMoME) and a novel More vs. Fewer (MoFe) ranking loss ensuring accuracy monotonicity: adding modalities should improve or maintain accuracy.

Methodology: DMoME consists of modality-specific expert networks (each an independent backbone) and a learnable gating network that dynamically adjusts each modality's contribution at the logit level (before softmax). The MoFe loss enforces L*MoFe = max(0, L*task(o*fewer, y) - L*task(o*more, y)), where o*more has more modalities than o_fewer. Two-stage training: (1) expert pretraining independently, minimizing interference; (2) cooperative learning jointly training experts + gating + MoFe loss. The logit-level weighting provides a temperature-rescaling effect for model calibration. Validated on BraTS 2018 (brain tumor segmentation, 4 MRI modalities, 15 modality configurations), UPMC Food-101 (image+text), and avMNIST (image+audio).

Main Evidence: On BraTS 2018 (15 modality configurations), SimMLM achieves highest average Dice scores across all settings. On UPMC Food-101, SimMLM achieves 72.20% image-only (vs. MoMKE 70.46%), 87.2% text-only (vs. MoMKE 86.59%), and 94.99% full modality (vs. MoMKE 92.71%). On avMNIST, 99.27% full modality. Calibration (ECE↓): SimMLM achieves 3.15%/3.75%/3.55% (ET/TC/WT) vs. MoMKE 3.46%/4.10%/4.04%. Confidence reliability (CR reduction): SimMLM achieves 50.68%/13.18%/61.97% CR reduction on ET/TC/WT vs. MoMKE. MoFe loss optimal coefficient λ=0.1. Expert pretraining is critical: skipping it causes significant performance drop. Logit-level mixture outperforms feature-level and probability-level mixture.

Relevance: The DMoME gating mechanism is architecturally analogous to the MM'25 paper's CBAM-inspired channel attention that concatenates shared and specific features before channel weighting — both adaptively weight modality contributions. The MoFe ranking loss could be applied to the MM'25 framework's training to ensure dual-modality accuracy ≥ single-modality accuracy, providing a principled training constraint. SimMLM's calibration results (ECE improvements of 0.31–0.35 percentage points) are particularly relevant to the clinical deployment gap identified in the MM'25 paper — uncertainty quantification is needed for trustworthy HER2 AI.

Limits: Validated on non-histopathology tasks (brain tumor segmentation, food classification, digit classification). Direct applicability to HER2 prediction on histopathology images is unproven. The MoFe loss requires training data with varying modality availability, which may not align with the paired H&E+IHC training setup.

Paper 8: MODES — Decoupled Multimodal Representation Fusion for Clinical Diagnostics (P008, npj Digital Medicine, 2025)

Problem and Task Setting: Simple multimodal concatenation causes semantic interference and information loss. MODES explicitly decouples shared features (common across modalities) from modality-specific features to improve fusion quality. Evaluated on a cardiovascular model using ECG and cardiac MRI (cMRI) from UK Biobank (4,150 held-out samples).

Methodology: Four components: (1) unimodal pre-trained encoders (foundation models) fine-tuned to encode shared (Z^s) and modality-specific (Z^n) representations; (2) unimodal generators that reconstruct original data from latent representations; (3) latent representations with masking component that eliminates low-information dimensions; (4) iterative three-step training: encode shared+specific → mask optimization → generator reconstruction. The masking component uses binary masks with L1 regularization controlled by β, enabling the model to infer the appropriate representation dimensionality. Comparison baselines: unimodal representations, early fusion, late fusion, DRIM (disentangled baseline).

Main Evidence: MODES representations outperform all baselines across a range of diagnostic phenotypes (RR interval, Ejection Fraction, etc.) and diagnoses (Atrial Fibrillation, valvular diseases). Missing modality: ECG representations can infer cMRI-related phenotypes, and vice versa, via the shared representation. Masking validation: predictive performance with masking is nearly identical to without masking, confirming compact representations without loss of information. The shared representation captures cross-modal diagnostic information while modality-specific representations encode unique information (ECG-specific: electrical activity; cMRI-specific: cardiac anatomy). Masking converges to different subspace sizes for different modality pairs depending on information content.

Relevance: Provides the strongest methodological validation for the MM'25 paper's pathological information decoupling encoder. MODES validates the shared-specific decomposition across a completely different domain (cardiovascular imaging), establishing it as a domain-agnostic principle. The MM'25 paper's DCO (Domain Classification Loss) and DAO (Distribution Alignment Loss) from Wang et al. CVPR 2023 represent a domain-specific instantiation of MODES-like decoupling in histopathology. The key insight that masking can reduce representation dimensionality without accuracy loss is applicable to the MM'25 paper's attention mechanism design.

Limits: Evaluated in cardiac imaging, not histopathology. The MM'25 paper's ResNet50-based shared encoder with DCO/DAO losses is a domain-specific instantiation. MODES requires pre-trained foundation models, which the MM'25 paper does not use.

Paper 9: CPath-Omni — Unified Multimodal Foundation Model for Computational Pathology (P009, arXiv, 2024)

Problem and Task Setting: Existing computational pathology models handle either patch-level or whole-slide image analysis, not both. CPath-Omni is a 15-billion-parameter multimodal LMM supporting both tasks, integrating pathology images with clinical text, trained with self-supervised objectives.

Methodology: Four-stage training: (1) patch-based pretraining aligning CPath-CLIP features with Qwen2.5-14B LLM using 700,145 image-caption pairs (CPath-PatchCaption); (2) fine-tuning with CPath-PatchInstruct for VQA, classification, captioning; (3) WSI pretraining with CPath-WSIInstruct using SlideParser for multi-scale WSI tokenization; (4) mixed patch-WSI training (15% random sampling) for knowledge transfer. CPath-CLIP combines OpenAI-CLIP-L and Virchow2 (DINOv2-based, 3M WSIs) as dual visual encoders, with Qwen2.5-14B as text encoder. SlideParser performs multi-scale region encoding (10×/20×/40×) and token compression (CoCa-style 1,152 query tokens) to standardize variable-sized WSI inputs.

Main Evidence: CPath-Omni achieves SOTA on 39/42 benchmarks. On PathMMU (largest pathology VQA dataset), CPath-Omni exceeds PathGen-LLaVA by 13.8% and surpasses human pathologist performance (71.8%) by 0.6%. On CPath-CLIP zero-shot classification, CPath-Omni surpasses PathGen-CLIP-L by 6.1%/7.7%/7.3% on Osteo/Pcam/LC-Lung. Few-shot: 95% accuracy on CRC with only 2 shots (vs. <91% for other models). On WSI-level tasks, CPath-Omni achieves performance comparable to task-specific ABMIL/DSMIL models. Training uses only 700K image-caption pairs, much fewer than general-domain CLIP models.

Relevance: Represents the computational pathology foundation model frontier. CPath-Omni's dual-visual-encoder design (CLIP + DINOv2) and its multi-scale WSI processing are architecturally distinct from the MM'25 paper's CNN-based approach, but the general finding that unified models can match task-specific models is encouraging for the long-term prospects of multimodal HER2 AI. CPath-Omni does not address missing modality robustness, suggesting this remains an open challenge even at the foundation model scale — validating the MM'25 paper's specific contribution.

Limits: Requires massive computational resources (15B parameters); missing modality handling is not specifically addressed; no HER2-specific evaluation reported; training from scratch vs. the MM'25 paper's fine-tuning scenario.

Paper 10: Staining Normalization Benchmarking with Multicenter Dataset (P010, Scientific Reports, 2026)

Problem and Task Setting: H&E stained tissue specimens exhibit significant color and staining variations across institutions, scanners, and staining protocols. This study benchmarks eight stain normalization methods on a unique multicenter dataset: the same tissue blocks (colon, kidney, skin) were stained at

66 different laboratories, isolating staining variation from other biological/technical variation.

Methodology: Four traditional methods (histogram matching, Macenko, Vahadane, Reinhard) and four deep learning methods (CycleGAN-UNet, CycleGAN-ResNet, Pix2pix-UNet, Pix2pix-DenseUNet). Quantitative evaluation: color transfer metrics (intersection, PCC, Euclidean distance, JS divergence in lαβ space), SSIM for structural similarity, FID for high-level feature similarity, Cellpose-SAM for nucleus detection counts, and foundation model (UNI-2) feature extraction with t-SNE visualization. WSI resampled to 10× for computational efficiency. Whole-WSI normalization (rather than patch-wise) applied for both traditional and deep learning methods to avoid tiling artifacts.

Main Evidence: No single method dominates across all tissue types and all metrics. Color transfer: histogram matching achieves best mean scores on skin (intersection=0.891, PCC=0.938, FID=61.67) and kidney (intersection=0.944, PCC=0.985). Colon: Macenko and Reinhard perform comparably. Structural similarity: Vahadane achieves highest SSIM (0.995) but worst color transfer; all methods maintain SSIM > 0.92. Nucleus detection: Pix2pix-UNet/CycleGAN-ResNet/CycleGAN-UNet yield counts closest to reference; histogram matching best among traditional methods. Foundation model features (t-SNE): CycleGAN and Macenko produce most compact clusters; Vahadane and Pix2pix-DenseUNet produce scattered distributions. Inference time: CycleGAN/Pix2pix ~4–5 min/WSI; histogram matching/Reinhard ~30s–2 min/WSI; Macenko/Vahadane ~2–7 min/WSI. Deep learning methods produce hallucination artifacts (CycleGAN-ResNet: false nuclei in adipose tissue; Pix2pix-DenseUNet: miscolored smooth muscle nuclei). Traditional methods: Macenko produces blue artifacts in eosin; Vahadane completely discards hematoxylin in many cases.

Relevance: Directly validates the single-institution limitation of the MM'25 paper's BCI dataset evaluation. The finding that no single stain normalization method dominates suggests that CM-GAN would need to be evaluated with stain normalization as a preprocessing step to assess robustness. The hallucination artifacts produced by deep learning-based normalization methods (CycleGAN, Pix2pix) are directly relevant to evaluating whether CM-GAN's cross-modal IHC generation artifacts could compromise downstream classification. The 66-laboratory dataset is freely available as a benchmark for future cross-institution HER2 AI validation.

Limits: Evaluated on H&E staining, not IHC. The MM'25 paper's CM-GAN generates IHC from H&E, which is a different task from H&E-to-standardized-H&E normalization. IHC-specific staining variation may differ from H&E variation in both magnitude and character.

Paper 11: Multi-Target Stain Normalization for Histology Slides (P011, arXiv, 2024)

Problem and Task Setting: Standard stain normalization uses a single reference image, which fails to capture the diversity of staining patterns in practical datasets. Multi-target normalization uses multiple reference images to improve robustness against the diversity of staining patterns encountered in multi-institutional settings.

Methodology: Parameter-free approach using multiple reference images for each target stain. Rather than computing stain statistics from a single canonical reference image, the method learns stain statistics from a diverse set of representative images that capture the range of staining variation observed in practice. The approach adapts to the statistical distribution of stain colors across the target dataset rather than forcing all images toward a single reference palette.

Main Evidence: Improves robustness against stain variation compared to single-reference approaches, with better generalization to external datasets. Specifically, the multi-reference approach avoids the pitfall of overfitting to the specific color distribution of a single reference image, which is the primary failure mode of single-reference methods when deployed on external data with different staining characteristics. The method does not require tuning of stain color parameters — it learns the appropriate normalization targets directly from the reference set.

Relevance: Directly addresses the cross-institution generalization concern raised by P010's finding that no single stain normalization method dominates. The MM'25 paper relies on CM-GAN to synthesize missing modalities but does not normalize the existing input modalities — multi-target normalization could serve as an input preprocessing step before CM-GAN, potentially reducing the distribution shift between H&E images from different institutions. The approach is particularly promising for the MM'25 paper because it could reduce the variance of CM-GAN's cross-modal reconstruction quality under staining variation, addressing the artifact exploitation hypothesis (H1) for the SSIM-accuracy dissociation. Unlike single-reference methods that produce poor results when the reference image is not representative of the target domain, multi-target normalization is inherently more robust to domain shift because it does not commit to a single stain color distribution.

Limits: arXiv preprint; limited empirical validation details. No quantitative comparison against single-reference methods in the available source. The specific quantitative improvement (e.g., "X% better color transfer than Macenko on external data") is not reported in the available snippet.

Paper 12: BCI — Breast Cancer Immunohistochemical Image Generation Through Pyramid Pix2pix (P012, CVPR 2022)

Problem and Task Setting: The foundational dataset and baseline method for the H&E-to-IHC cross-modal translation problem. The BCI dataset contains 4,870 registered H&E-IHC image pairs covering all four HER2 expression levels (0, 1+, 2+, 3+). The task is to synthesize IHC images from H&E images to reduce the cost and delay of actual IHC staining.

Methodology: Pyramid Pix2pix architecture: multi-scale residual blocks with skip connections across four Gaussian pyramid levels. Multi-scale loss: $L_{multi\text{-}scale} = \Sigma_i \lambda_i S_i$, combining GAN loss ($L$cGAN) and L1 pixel loss across scales. Dataset construction: unstained tissue → HE staining → IHC staining → elastix registration (16-block parallelized for efficiency) → image refinement → cutting. The registration quality is verified by overlap comparison with projection transformation.

Main Evidence: Pyramid Pix2pix outperforms standard Pix2pix, Pix2PixHD, and CycleGAN on the BCI benchmark. Gamma-corrected variant achieves PSNR=16.024 dB, SSIM=0.431 (weighted 0.6×SSIM + 0.4×PSNR ranking). Pix2pixHD incorrectly generates dark browns in low HER2 expression regions. Pyramid Pix2pix is better than CycleGAN and Pix2pix variants in both image quality and HER2 expression identification. Pathologist evaluation: 37.5% and 40.0% accuracy on generated IHC images by two pathologists — confirming that generated images are not yet clinically interpretable at the image level. The method is better at low HER2 expression (0/1+) than high expression (3+) in terms of generated image authenticity.

Relevance: The MM'25 paper's CM-GAN directly builds on BCI's Pyramid Pix2Pix as the foundation for its cross-modal module. The MM'25 paper's CM-GAN achieves PSNR=18.48 dB, SSIM=0.51, representing approximately an 18% relative SSIM improvement over BCI's gamma-corrected variant (SSIM=0.431), and a 15% improvement in PSNR. However, both remain relatively low SSIM values. Critically, the MM'25 paper's downstream classification accuracy (94.25% with Real H&E + Fake IHC) demonstrates that the classification head tolerates substantial reconstruction noise — the question is whether this tolerance generalizes to cross-institution deployment. The pathologist evaluation confirms that generated IHC images are not interpretable by human experts, limiting clinical utility of the generated images for direct diagnostic use. The 37.5–40% pathologist accuracy provides the specific baseline against which CM-GAN's improvement in human interpretability should be measured.

Limits: Evaluated only on image synthesis quality metrics, not on downstream HER2 classification accuracy from synthesized images. The BCI dataset is from a single institution, and the registration process (16-block elastix) may introduce artifacts in the paired dataset.

PDF-Refined Strengthened Analysis

The PDFs of the 10 downloaded papers provided substantial additional quantitative detail that strengthens the analyses above:

- **P001 (HER2 Interpretable Assessment)**: The PDF provides the complete per-class classification table (Table 5), showing that HER2 2+ is the most challenging class on HER2-IHC-40x-WSI (P=0.80, R=0.97), while HER2 3+ achieves perfect classification on HER2-IHC-40x-Patch. The MAP=79% score is quantified in the conclusion. The hybrid unfreezing ablation (Table 9) quantifies the frozen backbone's severe degradation on BCI (73% accuracy, Macro F1=0.69) vs. the proposed method (95% accuracy, Macro F1=0.95), providing concrete evidence for the magnification robustness challenge.
- **P006 (MoRA)**: The PDF provides the complete Tables 2-5 with all F1-Macro numbers across different modality missing configurations. The rank sensitivity analysis (Table 5) confirms r=4 is optimal (F1=80.73), and r=384 (full rank) gives 70.23 F1-Macro, worse than without MoRA, rigorously validating the low-rank hypothesis. The GPU memory and training time data (Table 3) quantifies the computational advantage.
- **P007 (SimMLM)**: The PDF provides the complete BraTS 2018 benchmark table (Table 1 with 15 modality configurations), the UPMC Food-101/avMNIST accuracy table (Table 2), and the calibration error table (Table 4). The logit-level vs. feature-level vs. probability-level mixture ablation (Figure A5) with concrete ECE numbers confirms the logit-level design choice. The MoFe ablation with λ sensitivity data provides the optimal λ=0.1 finding with quantitative support.
- **P008 (MODES)**: The PDF provides the four-component framework details (encoders, generators, representations, masking), the three-step iterative training procedure, and the quantitative phenotype prediction results. The shared-vs-specific representation analysis on the bMRI-cMRI modality pair confirms that some modality pairs have little shared information (e.g., electrical metrics from ECG vs. cardiac anatomy from cMRI), validating the decoupling approach in a different domain.
- **P009 (CPath-Omni)**: The PDF provides the complete four-stage training procedure, the CPath-CLIP architecture details (dual visual encoder combination), and quantitative results across 42 datasets. The human-level performance comparison (72.3% expert vs. CPath-Omni) provides the 0.6% surplus figure.
- **P010 (Stain Normalization)**: The PDF provides the complete quantitative results for all eight methods across three tissue types, the hallucination artifact descriptions, and the foundation model feature analysis with t-SNE. The inference time comparison quantifies the computational trade-offs between methods. Specific values cited: histogram matching intersection=0.891 on skin; CycleGAN FID=61.67; Vahadane SSIM=0.995.
- **P012 (BCI)**: The PDF provides the Pyramid Pix2pix architecture diagram (multi-scale pyramid with Gaussian convolution), the multi-scale loss formulation, the elastix registration pipeline with 16-block parallelization, and

the pathologist accuracy evaluation (37.5%/40.0%). The dataset statistics (3,123 WSI pairs, 1,750 patch pairs) and class distribution (Fig. 8) are extracted.

4.2 Integrated Thematic Assessment

4.2.1 Theme: The Missing Modality Problem and Why It Matters for HER2 Clinical Deployment

The clinical motivation for missing-modality robustness in HER2 prediction is compelling and well-supported by the literature. The systematic review (P003) documents that HER2-low (~40-50% of breast cancers) is the most therapeutically significant and diagnostically most difficult category, and that inter-observer variability at the 0/1+ boundary (60-75% concordance for 2+ cases) is a fundamental limitation of manual scoring. The existence of an automated approach that can predict HER2 expression from H&E alone (94.25% accuracy, +22.81% over baseline) would transform clinical workflows in regions with limited IHC staining infrastructure. The computational pathology review (P004) confirms that no existing HER2 AI system has been validated for HER2-low identification in routine clinical practice, highlighting both the clinical urgency and the current gap in the field.

The literature reveals three distinct architectural families for handling missing modalities in multimodal deep learning, each with different trade-offs relevant to the MM'25 paper's approach. The first family is **generative image synthesis**, exemplified by the CM-GAN and its BCI Pyramid Pix2Pix foundation (P012). The strength of this approach is that it produces visually interpretable synthesized images — even if the SSIM is low, pathologists can in principle examine the generated modality. The critical limitation is that BCI pathologists achieved only 37.5–40% accuracy interpreting generated IHC images (P012), confirming that low SSIM synthesis produces images that confuse even expert human observers. The second family is **modality-aware parameter-efficient adaptation**, exemplified by MoRA (P006), which uses modality-specific low-rank projections requiring only 1.6% trainable parameters. This approach is far more computationally efficient than generative synthesis and achieves strong results on image+text modalities, but requires pre-trained multimodal transformers and has not been demonstrated on histopathology image+image tasks. The third family is **dynamic expert gating**, exemplified by SimMLM's DMoME (P007), which uses logit-level mixture of modality experts with a ranking loss enforcing accuracy monotonicity. This approach is architecturally closest to the MM'25 paper's CBAM-inspired attention module, and its calibration improvements (ECE reductions of 0.31–0.35 percentage points on BraTS) are directly relevant to the uncertainty quantification gap.

The structural motivation for choosing generative synthesis over feature/token-level adaptation in the H&E+IHC domain is rooted in the nature of the cross-modal translation task. H&E-to-IHC translation involves pixel-level semantic mapping — converting morphological tissue structures visible in H&E (nuclear

shape, glandular architecture, stromal patterns) into membrane protein expression patterns highlighted by IHC — which is fundamentally a pixel-level image-to-image translation problem rather than a representation-level alignment problem. In contrast, MoRA's modality-aware projections and SimMLM's gating operate at the feature/token level; when applied to image+image modality pairs, they must bridge a larger semantic gap without producing a direct pixel-level reconstruction, making it harder to verify whether the adapted representations capture the precise subcellular staining patterns that distinguish HER2 expression levels. Generative synthesis, despite its lower SSIM, directly outputs a full-resolution IHC image that the downstream classifier can leverage — the pixel-level approach is structurally more aligned with the domain's information structure, even if the synthesis quality is imperfect.

The strongest evidence supports the conclusion that the MM'25 paper's generative approach is a meaningful contribution but may not be the most parameter-efficient solution for missing-modality robustness. A comparative evaluation of CM-GAN vs. MoRA-style adaptation vs. SimMLM-style gating on the same HER2 classification task would be the definitive experiment, and the literature suggests that multiple approaches may coexist — generative synthesis for interpretability, and parameter-efficient adaptation for efficiency.

4.2.2 Theme: Feature Decoupling — Validated Architecture Across Domains

The pathological information decoupling encoder in the MM'25 paper — which separates shared features (common morphological patterns in H&E and IHC) from modality-specific features (structural patterns in H&E; protein expression patterns in IHC) — is one of the most architecturally significant contributions, and it is independently validated by MODES (P008, npj Digital Medicine 2025).

Why shared-specific decoupling helps: The complementary information hypothesis. The mechanistic rationale for decoupling shared and specific features rests on the fact that H&E and IHC images encode partially overlapping but non-identical biological information. Shared features capture what both modalities have in common: the overall tissue architecture (glandular structure, stromal distribution, nuclear density), which reflects the underlying tumor morphology that determines both H&E appearance and IHC staining patterns. These shared features are the basis for cross-modal transfer — they are why H&E→IHC synthesis is feasible at all, because there exists a common morphological substrate visible in both stains. Specific features capture what each modality reveals uniquely: H&E-specific features encode tissue texture, nuclear morphology, architectural atypia; IHC-specific features encode membrane protein expression intensity, the spatial distribution of positively stained cells, and the membrane completeness scoring criteria that pathologists use for HER2 grading. By separating these two streams, the decoupling encoder prevents interference: IHC-specific membrane features are not diluted by H&E texture patterns in the shared pathway, and H&E-specific structural features do not contaminate the IHC-specific protein expression pathway. MODES (P008)

validates this principle in a completely different domain — ECG and cardiac MRI share cardiac function information (heart rate variability, ejection fraction) while each encodes unique information (electrical activity vs. anatomical structure) — confirming that the complementary information hypothesis is domain-agnostic rather than histopathology-specific.

The MM'25 paper's instantiation using DCO (Domain Classification Loss) and DAO (Distribution Alignment Loss) from Wang et al. CVPR 2023 operationalizes this principle through a contrastive training objective: DCO pushes the shared and specific feature distributions apart by training a domain discriminator to distinguish between them, while DAO pulls the distribution of shared features from H&E and IHC into alignment so that they are comparable representations of the same underlying morphology. The CBAM-inspired channel attention then learns to weight the shared and specific contributions based on input quality — when the real IHC modality is available, the IHC-specific pathway contributes high-quality membrane information and the attention module upweights it; when only synthetic IHC is available (H&E→IHC reconstruction), the attention module downweights the specific pathway and relies more heavily on the shared morphological representation. This adaptive weighting is why the attention module contributes +4.91% F1 improvement: it is essentially a learned gating mechanism that selects the most reliable information stream for each input, rather than using a fixed fusion rule.

MODES also provides a key architectural insight: the masking component (which eliminates low-information representation dimensions) maintains accuracy while reducing representation dimensionality. The MM'25 paper's attention mechanism — which concatenates shared and specific features before applying channel attention — is conceptually consistent with this finding: the channel attention is effectively learning to mask low-value feature channels. A natural follow-up would be to explicitly incorporate MODES-style masking into the MM'25 framework, which could improve representation compactness and potentially generalization.

The strongest evidence here is the MODES result that shared representations can infer modality-related phenotypes even when the source modality is missing — this is precisely analogous to the MM'25 paper's ability to predict HER2 from H&E alone using the shared representation pathway, confirming that the shared feature stream is the primary mechanism of missing-modality robustness.

4.2.3 Theme: Cross-Modal Reconstruction Quality and Its Relationship to Classification Accuracy

The most counterintuitive finding in the literature is the dissociation between cross-modal reconstruction quality (SSIM = 0.39–0.51) and downstream classification accuracy (94.25%). This dissociation is important to understand because it has direct implications for clinical deployment.

Why CNNs tolerate low-SSIM synthesis: The feature extraction hypothesis. The BCI paper (P012) provides the key evidence with specific quantification: pathologists achieved only 37.5% and 40.0% accuracy interpreting generated IHC images by two independent observers — confirming that the generated images are visually misleading at the pixel level even when they contain sufficient discriminative information for a CNN classifier. The SSIM of generated images (0.431 for BCI, 0.51 for CM-GAN) is far below the SSIM > 0.90 range typically considered acceptable for natural image synthesis, and the pathologist accuracy confirms that these low-SSIM images are genuinely confusing to human experts, not merely aesthetically imperfect.

The mechanistic explanation for why CNNs nevertheless maintain high accuracy is rooted in the nature of CNN feature extraction: the downstream ResNet50 classification head operates on high-level convolutional feature maps rather than pixel-level image content. CNNs learn hierarchical representations where early layers capture texture and edge patterns while deeper layers capture abstract semantic features (nuclear morphology, glandular architecture, membrane staining intensity). The H&E→IHC cross-modal reconstruction captures sufficient structural information — the approximate shape of membrane regions, the relative density of positively vs. negatively stained cells, the overall tissue architecture — for the CNN feature extractor to classify HER2 expression levels, even when pixel-level details (exact membrane brown staining intensity, precise subcellular localization) are incorrectly rendered. This is analogous to why JPEG-compressed images at quality 30 still yield near-identical classification accuracy to originals: CNNs are robust to distributional noise at the level of visual perception because their decision boundary is carved in a feature space that smooths out pixel-level artifacts. Critically, this tolerance is a feature of the CNN architecture, not a guarantee of robustness under domain shift — the question is whether this same tolerance holds when the artifact patterns change (cross-institution deployment), which is why BCI pathologist evidence confirms the artifact problem on single-institution data but does not validate cross-institution robustness.

Why cross-institution variation is the critical threat: The artifact exploitation hypothesis. The stain normalization benchmarking (P010) adds another analytical layer with specific quantitative evidence: even deep learning-based stain normalization methods (CycleGAN, Pix2pix) produce hallucination artifacts on histopathology images — CycleGAN-ResNet generates false nuclei in adipose tissue; Pix2pix-DenseUNet produces miscolored smooth muscle nuclei. Traditional methods produce different artifacts: Macenko produces blue artifacts in eosin; Vahadane completely discards hematoxylin in many cases. The color transfer data from P010 quantifies the magnitude: histogram matching achieves intersection=0.891 on skin but CycleGAN's FID=61.67 on the same tissue, indicating substantial structural distortion. Critically, no single normalization method dominates across all tissue types and all metrics, meaning CM-GAN's

cross-modal synthesis may similarly hallucinate staining patterns in tissue regions where H&E and IHC staining have less predictable relationships.

The mechanism for why this matters is two-fold. First, when CM-GAN is trained on the BCI single-institution dataset, it learns not only the genuine morphological relationship between H&E tissue structures and IHC protein expression, but also the institutional quirks: the specific scanner characteristics, the particular staining protocol's color profile, the tissue preparation artifacts that are consistent within one institution but absent elsewhere. The shared representation learned by the pathological information decoupling encoder captures both genuine cross-modal morphology AND these institution-specific patterns. Second, when deployed at a new institution, the H&E images from the new institution have different scanner characteristics and staining chemistry — the shared encoder's response to these different inputs is unknown. If the new institution's tissue artifacts happen to resemble the BCI training artifacts, cross-institution accuracy may remain high (H1a). If they differ substantially, the shared representation's learned features may not activate correctly, and the classification head may be unable to compensate (H1b). P010's evidence of dramatic visual differences across 66 laboratories on the same tissue block suggests that H1b is the more likely outcome for models trained on single-institution data — the "artifact exploitation" concern is not hypothetical but empirically grounded in the magnitude of cross-institution staining variation.

The multi-target stain normalization approach (P011) offers a potential mitigation: using multiple reference images for normalization rather than a single canonical reference, which could reduce the hallucination artifact rate by avoiding overfitting to a single stain color distribution. Integrating multi-target normalization as a preprocessing step before CM-GAN could be a valuable experiment specifically for addressing the artifact exploitation hypothesis.

4.2.4 Theme: Cross-Institution Generalization — The Critical Validity Threat

The single-institution evaluation of the MM'25 paper on the BCI dataset is the most significant validity threat, and the stain normalization literature (P010) provides the most concrete evidence of why this matters. The 66-laboratory multicenter study demonstrates that even the same tissue block stained at 66 different labs produces dramatic visual differences: histogram matching achieves intersection scores ranging from 0.891 (skin) to 0.944 (kidney) against a reference stain, while FID scores range from 61.67 (histogram matching on skin) to much higher values for deep learning methods. Critically, no single stain normalization method dominates across all tissue types and all metrics — histogram matching is best for skin and kidney, while Macenko and Reinhard perform comparably for colon. This means that a model trained on data from one institution's staining protocol faces a fundamentally different input distribution when deployed at another institution, even if the underlying tissue biology is identical.

The ResViT-GANNet paper (P005) reinforces this concern with specific numbers: its dual-branch multimodal architecture (CNN + ViT) achieves 96.40% accuracy on BreakHis, demonstrating that multimodal histopathology classification is broadly feasible — but BreakHis itself is a multi-institutional dataset, and the variation in accuracy across magnifications and tumor types (8 tumor types × 4 magnifications) suggests sensitivity to data distribution. The 3.3% accuracy improvement from GAN augmentation specifically on minority tumor classes indicates that data diversity matters for classification performance on difficult cases.

The HER2 MRI multimodal paper (P002) provides an interesting counterpoint: ResNet50 + clinical nomogram achieves AUC=0.94 using late fusion of MRI features with clinical metadata, demonstrating that HER2 prediction from non-histopathology imaging is feasible and that incorporating clinical variables adds discriminative value. This suggests a potential extension for the MM'25 framework — combining histopathology features with clinical metadata (tumor grade, Ki-67 index, hormone receptor status) could improve robustness under cross-institution variation.

The direct implication for the MM'25 paper is that the 94.25% H&E-only accuracy should be validated on external datasets with different staining protocols before clinical deployment. The BCI dataset's single-institution origin means that the reported performance may be optimistic for cross-institution deployment.

4.2.5 Theme: Interpretability and Uncertainty — The Clinical Deployment Gaps

Two critical clinical deployment requirements are identified across the literature but not addressed by the MM'25 paper: interpretability and uncertainty quantification. These gaps are interconnected — without calibrated confidence estimates, clinicians cannot reliably distinguish between high-confidence correct predictions and uncertain predictions that require human review; and without interpretable attention maps, even confident predictions cannot be explained to pathologists or patients.

Interpretability: The attention mechanism lacks membrane-level quantitative validation. The HER2 interpretable assessment paper (P001) establishes MAP (Membrane Activation Precision) and EC (Explanation Consistency) as concrete quantitative metrics for HER2 IHC scoring interpretability, evaluated against pathologist-annotated membrane masks. The MAP = 79% score for the hybrid unfreezing ResNet50 indicates that 79% of the model's attention overlaps with pathologist-annotated membrane regions — meaning roughly one-fifth of the model's focus falls on tissue regions that pathologists do not consider morphologically relevant to HER2 scoring. The MM'25 paper's CBAM-inspired attention module is architecturally positioned to produce similar interpretability maps, concatenating shared and specific features before channel-wise attention, which is conceptually consistent with the attention mechanism in P001's Score-CAM pipeline. However, no MAP/EC-style evaluation

is reported in the MM'25 paper — the attention weights are visualized but not validated against pathologist-annotated ground truth. This matters clinically because MAP = 79% (P001) is acceptable but not optimal; improvements in attention alignment could directly translate to more clinically trustworthy explanations. Furthermore, P001's WSI-level results show substantially lower HER2 1+ precision (P=0.82, R=0.83) than patch-level results on BCI (P=0.95, R=0.90, F1=0.85), suggesting that attention quality degrades at the WSI scale — the MM'25 paper's evaluation is patch-level on BCI, so its attention module quality at WSI scale is unknown.

Uncertainty quantification: No calibration, no clinical trust. SimMLM (P007) provides the strongest quantitative evidence for calibrated uncertainty in multimodal deep learning: the logit-level DMoME gating mechanism achieves ECE = 3.15%/3.75%/3.55% (ET/TC/WT) on BraTS 2018, compared to 3.46%/4.10%/4.04% for the baseline MoMKE method — a reduction of 0.31–0.35 percentage points in calibration error. Crucially, SimMLM achieves confidence reliability (CR) reduction of 50.68%/13.18%/61.97% across the three MRI sequences, meaning the model's confidence estimates are substantially more predictive of actual correctness. This is the specific mechanism by which uncertainty quantification translates to clinical utility: not just lower ECE, but higher confidence reliability. The MM'25 paper provides no confidence estimates — a critical gap for clinical deployment. P003 and P004 explicitly identify the absence of uncertainty quantification as one of the three major barriers to clinical deployment for HER2 AI systems, alongside stain variation and limited data diversity. The clinical implication is severe: on HER2 1+ cases, where inter-observer concordance is lowest (60–75% for 2+ cases per P003, substantially lower at the 0/1+ boundary), the model's predictions are most uncertain but most therapeutically consequential — T-DXd eligibility depends on correct HER2-low classification. Without calibrated confidence estimates, clinicians cannot systematically identify which HER2 1+ predictions require pathologist escalation. The concrete calibration target for the MM'25 framework is ECE < 5% per P007's evidence from BraTS, with per-class ECE < 8% for HER2 1+ specifically (since the HER2 1+ category is the most uncertain and the highest therapeutic stakes).

The interconnection between interpretability and uncertainty. P001 demonstrates that Expected Calibration Error (ECE) is part of the interpretability evaluation framework — the evaluation framework includes ECE to assess model confidence reliability alongside MAP and EC. This means that improving uncertainty quantification and improving interpretability attention maps are not independent goals; they share the same calibration quality: a model with well-calibrated confidence (low ECE) and high MAP produces predictions that are both explainable and trustworthy. The MM'25 paper's CBAM-inspired attention module is the most natural integration point for both goals — if the channel attention can be trained with a calibration-aware loss (following SimMLM's MoFe principle with $\lambda=0.1$) and evaluated with MAP against pathologist-annotated membrane

regions, it could simultaneously address both the interpretability and uncertainty quantification gaps identified in this theme.

The foundation model context. CPath-Omni (P009) suggests a longer-term path: a 15B-parameter foundation model achieving SOTA across 42 computational pathology benchmarks, approaching human pathologist performance (72.3% expert vs. CPath-Omni 71.8% on PathMMU with a 0.6% surplus). However, CPath-Omni does not address missing modality robustness, indicating that this specific capability remains an open research problem even at the foundation model scale — validating the MM'25 paper's specific contribution. The absence of uncertainty quantification in CPath-Omni also suggests that this remains an open challenge at the frontier of computational pathology.

5. Integrated Assessment for the Current Project

The evidence from the grounded source and the literature supports the following integrated assessment for this HER2 prediction framework:

Strongly justified directions: The dual-modality flexible architecture is well-motivated by a genuine clinical need (missing modality in resource-limited settings) and is supported by convergent evidence from three independent research directions (generative synthesis per P012, feature decoupling per P008, dynamic gating per P007). The feature decoupling approach is independently validated by MODES (P008, npj Digital Medicine 2025) across a different clinical domain (cardiovascular imaging with ECG and cMRI), substantially strengthening confidence in the architectural design. The CM-GAN's 22.81% improvement over the H&E unimodal baseline confirms that cross-modal reconstruction adds substantial discriminative value, not merely cosmetic enhancement. The CBAM-inspired attention module's +4.91% F1 improvement is consistent with SimMLM's findings on logit-level modality weighting (P007). The 96.40% accuracy achieved by ResViT-GANNet (P005) on multi-institutional BreakHis data confirms that the ~95% accuracy range is achievable under real-world multi-institutional conditions.

Why the SSIM-accuracy dissociation holds on BCI but may not generalize. The mechanistic explanation for why CM-GAN achieves 94.25% accuracy despite SSIM=0.51 rests on the CNN feature extraction hypothesis: the ResNet50 classification head operates on high-level convolutional features that tolerate pixel-level synthesis artifacts because the features relevant for HER2 classification (overall membrane region density, nuclear-to-cytoplasmic ratio patterns, glandular architecture) are preserved even when precise subcellular staining details are incorrectly rendered. This is analogous to why CNNs are robust to JPEG compression at quality 30 — feature-space decisions are insensitive to pixel-level noise that human observers find visually confusing. BCI pathologists' 37.5-40% accuracy interpreting generated IHC images (P012) confirms that the SSIM=0.51 images are genuinely misleading to human perception, establishing a clear human-machine performance gap on the same visual inputs. However, this tolerance may be specific to the BCI single-institution

setting where the classification head has learned to rely on institution-specific artifact patterns alongside genuine morphological features. Cross-institution deployment could expose whether the robustness generalizes (H2: genuine shared-feature robustness) or collapses (H1: artifact exploitation specific to BCI's training distribution).

Why cross-institution variation is the critical threat. The mechanism for why cross-institution deployment is the most serious validity threat has three components. First, when CM-GAN is trained on BCI's single institution, it learns not only the genuine morphological relationship between H&E tissue structures and IHC protein expression but also the institutional quirks: specific scanner characteristics, particular staining chemistry, tissue preparation artifacts that are consistent within one institution but absent elsewhere. Second, the pathological information decoupling encoder's shared features capture both genuine cross-modal morphology AND these institution-specific patterns — DCO pushes shared and specific features apart, but institution-specific features are shared across the training institution's H&E and IHC images, so they leak into the shared representation pathway. Third, P010's evidence from 66 laboratories on the same tissue block demonstrates that the same biological tissue produces dramatically different visual outputs across institutions: histogram matching achieves color transfer intersection scores of only 0.891-0.944 even with the best methods, meaning 6-11% of stain color information is unaligned. When CM-GAN encounters H&E images from a new institution with different staining chemistry, the shared encoder's response is unpredictable — if the new institution's tissue artifacts resemble BCI training artifacts, accuracy may remain high (H1a); if they differ substantially, the shared representation may activate incorrectly and cross-modal synthesis quality will degrade (H1b). The magnitude of P010's cross-institution variation makes H1b the more likely outcome for real-world deployment, making the artifact exploitation hypothesis the most critical threat to clinical validity.

Cross-cutting synthesis across literature strands: The three architectural families for missing-modality robustness (generative synthesis per P012, parameter-efficient adaptation per P006, dynamic gating per P007) are not mutually exclusive — they address different aspects of the problem. CM-GAN prioritizes visual interpretability of synthesized modalities; MoRA prioritizes computational efficiency; SimMLM prioritizes calibration. The MM'25 paper's contribution is strongest at the intersection of these: it demonstrates that generative synthesis of cross-modal features (CM-GAN) can be combined with shared-specific decoupling (DCO/DAO) and attention-based fusion (CBAM) in a single end-to-end pipeline. This combination is novel and is not found in any single referenced paper.

How MODES and SimMLM jointly constrain the design space: MODES (P008) establishes that shared-specific decoupling is necessary for high-quality multimodal fusion — concatenation-based approaches leave accuracy on the table (+4.29% improvement from decoupling over concatenation in the MM'25

paper). SimMLM (P007) establishes that calibration is achievable through logit-level modality weighting (ECE reductions of 0.31–0.35 percentage points). Jointly, these findings suggest that the MM'25 framework's ideal training objective should: (1) enforce shared-specific decoupling via DCO/DAO (as it currently does), (2) add a calibration-aware loss term (following SimMLM's MoFe principle with λ=0.1), and (3) evaluate uncertainty using ECE metrics. The existing CBAM attention module is architecturally well-positioned for the calibration extension.

Weakly supported assumptions: The framework's cross-institution generalization is entirely unvalidated. The 66-laboratory stain normalization evidence (P010) — demonstrating that even the same tissue block stained at 66 different labs produces intersection scores ranging from 0.891 to 0.944 depending on tissue type and method — suggests that the performance reported on the BCI dataset may not transfer to other institutions. The CM-GAN's generative approach may produce hallucination artifacts analogous to those documented for CycleGAN and Pix2pix in P010 (false nuclei in adipose tissue, miscolored smooth muscle nuclei). The HER2 1+ performance is unknown — aggregate accuracy of 94.25% may mask degraded performance at the HER2-low boundary, which is the most clinically important and most difficult category per P003.

Assumptions that should be treated cautiously: The high classification accuracy despite low reconstruction SSIM should not be interpreted as evidence that reconstruction quality is unimportant. The classification head may be exploiting dataset-specific artifacts (BCI registration patterns, institutional staining quirks) rather than genuine HER2-relevant morphological features, in which case cross-institution deployment could see a larger accuracy drop than the single-institution evaluation suggests. The BCI pathologist evidence (37.5–40% accuracy interpreting generated images, P012) confirms that generated images are not human-interpretable at the current SSIM level — this limits the framework's clinical utility in human-AI hybrid workflows. The fact that CM-GAN achieves 94.25% accuracy with fake IHC does not mean the fake IHC images are medically meaningful; it means the classifier tolerates their imperfections.

What the current evidence does not allow us to conclude: We cannot conclude that the framework is ready for clinical deployment without external validation on multi-institutional data. We cannot conclude that CM-GAN is superior to MoRA-style or SimMLM-style missing-modality approaches without direct comparative benchmarking on the same HER2 classification task. We cannot conclude that the framework handles the HER2 0/1+ boundary robustly without per-class accuracy reporting. We cannot conclude whether the SSIM-accuracy dissociation reflects genuine shared-feature robustness (H2) or dataset-specific artifact exploitation (H1) without cross-institution evidence.

6. Unresolved Questions and Decision-Critical Gaps

1. **Cross-institution generalization**: The framework is evaluated exclusively on the BCI dataset from a single institution. The 66-laboratory stain

normalization study (P010) demonstrates that staining variation across institutions is substantial and cannot be resolved by a single normalization method. The color transfer intersection scores range from 0.891 (histogram matching on skin) to 0.944 (histogram matching on kidney), indicating that even the best methods leave ~5-10% of color information unaligned across institutions. An external validation experiment on datasets from at least 2–3 different institutions with different scanners and staining protocols is essential before clinical deployment considerations.

2. **HER2 1+ / HER2-low boundary performance**: The most therapeutically important and diagnostically most difficult category is evaluated only implicitly. The aggregate 94.25% accuracy may mask substantially degraded performance at the HER2-low boundary, as demonstrated by the P001 paper's finding of F1 = 0.85 for HER2 1+ on BCI 10× (vs. F1=0.99 for HER2 3+), and P=0.82, R=0.83 on HER2-IHC-40x-WSI for Class 1+. Per-class confusion matrix analysis is needed. The inter-observer concordance of 60-75% for 2+ cases per P003 underscores that the boundary is genuinely difficult even for human experts.

3. **Uncertainty quantification**: The framework provides no confidence estimates, which is a near-essential requirement for clinical decision support. Clinicians need to know not just the predicted class but the probability that the prediction is correct, especially for borderline cases. SimMLM (P007) provides a concrete calibration target: ECE < 5% on the BCI dataset. Monte Carlo dropout or temperature scaling could produce calibrated probabilities; the correction rate on uncertain predictions (<80% confidence) when reviewed by a pathologist could serve as the validation metric.

4. **Interpretability evaluation**: The CBAM-inspired attention module's output is not evaluated against pathologist-annotated membrane regions using metrics like MAP and EC. Without this evaluation, it is unclear whether the attention mechanism produces clinically meaningful explanations. P001 provides the concrete targets: MAP > 75%, EC > 80%. Applying Score-CAM to the attention module output and comparing attention regions against pathologist-annotated membrane masks would address this gap.

5. **CM-GAN vs. alternative missing-modality approaches**: Three distinct architectural families have been validated independently but never benchmarked against each other on the same HER2 classification task. It is possible that MoRA-style or SimMLM-style approaches achieve comparable or superior missing-modality robustness with lower computational cost (MoRA uses only 1.6% trainable parameters per P006). Comparative benchmarking on the BCI dataset would determine whether the computational cost of generative synthesis is justified.

6. **Relationship between reconstruction quality and cross-institution robustness**: The MM'25 paper's high classification accuracy despite low

reconstruction SSIM may be specific to the single-institution BCI dataset. If the classification head exploits dataset-specific artifacts, cross-institution deployment could see accuracy drop substantially. P010 provides the mechanism: CycleGAN and Pix2pix produce hallucination artifacts (false nuclei in adipose tissue, miscolored smooth muscle nuclei), and CM-GAN may produce analogous artifacts under cross-institution staining variation. Correlation analysis between reconstruction SSIM and classification accuracy drop under domain shift would identify whether synthesis quality predicts cross-institution robustness.

7. **Generalizability to other immunohistochemical biomarkers**: The framework is validated only for HER2. The HER2 MRI paper (P002) demonstrates that multimodal HER2 prediction from different imaging modalities (MRI + clinical nomogram) achieves AUC=0.94, suggesting that the multimodal fusion approach generalizes across imaging modalities. However, it is unclear whether the CM-GAN cross-modal synthesis specifically generalizes to other biomarker prediction tasks (ER/PR, Ki-67) or whether HER2-specific characteristics (strong membrane staining, well-defined expression levels) make it uniquely suited to this approach.

8. **Clinical workflow integration**: The framework assumes availability of WSI-level inputs but does not address the practical challenge of region-of-interest (ROI) selection, tissue segmentation, or handling of out-of-tissue regions in whole-slide images. P001 provides evidence that the frozen backbone approach degrades severely from 95% to 73% accuracy on BCI at 10×, suggesting that backbone selection and fine-tuning strategy matter for robustness. End-to-end evaluation with automatic tissue segmentation and ROI detection would address this gap.

7. Recommended Next Steps

All recommendations below are prioritized by clinical impact and grounded in specific evidence from the literature. Each recommendation is structured with: (1) the specific action to take; (2) the metric to measure; (3) the target value to achieve; and (4) the baseline for comparison.

1. **Conduct cross-institution external validation** (highest priority, targeted at single-institution validity threat per P010):

- **What to do**: Train the CM-GAN framework on the BCI single-institution dataset; evaluate on 2–3 external HER2 datasets from institutions with different scanners, staining protocols, and tissue preparation methods (e.g., available HER2 datasets from other hospitals or the 66-laboratory benchmark from P010). - **Metric to measure**: Per-institution accuracy, accuracy drop from single-institution to multi-institution average. - **Target to achieve**: Accuracy drop < 5 percentage points from the BCI baseline (94.25% H&E-only; 95.09% dual-modality). - **Baseline to compare against**: BCI single-institution 94.25% (H&E-only), 95.09% (dual-

modality) per source paper Table 1. Use P010's 66-laboratory intersection scores (0.891–0.944) as reference for expected staining variation magnitude.

2. **Report per-class HER2 performance with confusion matrix** (targeted at HER2-low evaluation gap per P001, P003):

- **What to do**: Add stratified per-class precision, recall, F1, AUC, and a full 4×4 confusion matrix to the evaluation framework, using class-balanced sampling to ensure adequate HER2 1+ representation. - **Metric to measure**: Per-class F1-score for each HER2 class (0, 1+, 2+, 3+). - **Target to achieve**: HER2 0 F1 ≥ 0.95; HER2 1+ F1 ≥ 0.85 (matching P001's F1=0.85 on BCI 10×); HER2 2+ F1 ≥ 0.90; HER2 3+ F1 ≥ 0.95. - **Baseline to compare against**: P001's per-class BCI results at 10× magnification — Class 0 (P=0.97, R=0.97), Class 1+ (P=0.95, R=0.90, F1=0.85), Class 2+ (P=0.94, R=0.98), Class 3+ (P=0.97, R=0.94).

3. **Integrate uncertainty quantification with calibrated probability output** (targeted at clinical deployment gap per P003, P004, P007):

- **What to do**: Implement Monte Carlo dropout with 50+ forward passes (dropout enabled at inference) to produce ensemble-based confidence estimates; apply temperature scaling post-training to calibrate the probability outputs; generate reliability diagrams and confidence histograms per HER2 class. - **Metric to measure**: Expected Calibration Error (ECE) on the BCI validation set, broken down per HER2 class. - **Target to achieve**: ECE < 5% overall; per-class ECE < 8% for HER2 1+ (the most uncertain category). - **Baseline to compare against**: SimMLM (P007) achieves ECE = 3.15%/3.75%/3.55% (ET/TC/WT) on BraTS, with confidence reliability reduction of 50.68%/13.18%/61.97%. Temperature scaling on the MM'25 framework should target the SimMLM calibration level as a reasonable clinical standard.

4. **Implement and benchmark MoRA-style and SimMLM-style missing-modality alternatives on BCI** (targeted at comparative architecture question per P006, P007):

- **What to do**: Implement MoRA-style modality-aware LoRA adaptation (rank r=4, 1.6% trainable parameters per P006) and SimMLM-style DMoME gating (MoFe loss coefficient λ=0.1 per P007) on the BCI HER2 classification task; train all three methods (CM-GAN, MoRA, SimMLM) on the same BCI train/val/test split; evaluate H&E-only, IHC-only, and dual-modality accuracy for all three. - **Metric to measure**: H&E-only accuracy, IHC-only accuracy, dual-modality accuracy, per-class F1, GPU memory usage (GB), training time per 1,000 steps (hours). - **Target to achieve**: If MoRA or SimMLM achieves within 2 percentage points of CM-GAN's accuracy on the same test split, the parameter-efficient approach is preferred for resource-limited deployment; report the accuracy-memory trade-off explicitly. - **Baseline to compare against**: CM-GAN's current results — 94.25% H&E-only, 90.28% IHC-only, 95.09% dual-modality per source paper Table 1. MoRA's ODIR result: F1-Macro=92.56 at 30% image + 100% text missing per

P006 Table 2. SimMLM's UPMC Food-101 result: 72.20% image-only, 87.2% text-only per P007 Table 2.

5. **Evaluate CBAM attention interpretability using MAP/EC metrics with pathologist validation** (targeted at interpretability gap per P001):

- **What to do**: Apply Score-CAM to the CBAM-inspired attention module's output; compute MAP (Membrane Activation Precision) as the overlap between model attention regions and pathologist-annotated membrane regions; compute EC (Explanation Consistency) as the stability of attention regions across similar HER2 expression inputs; have ≥ 2 board-certified pathologists annotate membrane regions on a random subset of 200 BCI images covering all four HER2 classes. - **Metric to measure**: MAP (percentage of model attention overlapping with pathologist-annotated membrane regions); EC (percentage of attention regions that remain stable across ≥ 3 similar inputs per P001 definition). - **Target to achieve**: MAP > 75% (matching P001's MAP=79% for hybrid unfreezing ResNet50); EC > 80%. - **Baseline to compare against**: P001's ResNet50 hybrid unfreezing achieves MAP=79% on BCI at 10×, with EC quantified in the conclusion. The frozen backbone MAP was substantially lower, establishing that unfreezing is critical for interpretability.

6. **Apply multi-target stain normalization as preprocessing before CM-GAN cross-modal synthesis** (targeted at hallucination artifact concern per P010, P011):

- **What to do**: Integrate multi-target stain normalization (P011) as an input preprocessing step before CM-GAN's H&E→IHC synthesis; test both Macenko (best FID=61.67 for kidney per P010) and histogram matching (best intersection=0.891 for skin per P010) as alternative preprocessing strategies; use the 66-laboratory stain normalization benchmark dataset (P010) to evaluate reconstruction artifact rate under multi-target normalization vs. no normalization. - **Metric to measure**: Reconstruction artifact rate (percentage of images with visible hallucination artifacts per pathologist evaluation, analogous to P010's false nuclei and miscolored nuclei assessment); reconstruction SSIM and PSNR on external data. - **Target to achieve**: Reduction in reconstruction artifact rate > 20% on external datasets (different institution staining) compared to unnormalized inputs; maintain or improve reconstruction SSIM ≥ 0.45 on external data. - **Baseline to compare against**: P010 documents that CycleGAN-ResNet generates false nuclei in adipose tissue and Pix2pix-DenseUNet produces miscolored smooth muscle nuclei under cross-institution variation. P010's histogram matching achieves intersection=0.891 on skin and intersection=0.944 on kidney — use these as reference scores for stain normalization effectiveness.

7. **Investigate HER2 0/1+ binary boundary performance using hybrid unfreezing** (targeted at HER2-low boundary gap per P001, P003):

- **What to do**: Conduct a focused binary HER2 classification evaluation on HER2 0 vs. HER2 1+ cases (the most therapeutically important boundary per P003 — T-DXd eligibility); apply the hybrid unfreezing training strategy from P001 to the MM'25 paper's IHC unimodal branch (freeze ResNet50 Blocks 1–2 for generic histological features, fine-tune Blocks 3–4 and fc); use class-balanced sampling to ensure equal HER2 0 and HER2 1+ representation in training and evaluation. - **Metric to measure**: Binary HER2 0/1+ AUC-ROC; binary accuracy; per-class precision and recall for the 0/1+ boundary specifically. - **Target to achieve**: HER2 0/1+ binary AUC ≥ 0.92 on BCI (aligning with the Corr-A-Net benchmark of AUC=0.98 for H&E→HER2 from search snippet); IHC unimodal branch accuracy ≥ 90%. - **Baseline to compare against**: P001's BCI 10× HER2 1+ results — F1=0.85, P=0.95, R=0.90; P003's documented inter-observer concordance of 60–75% for equivocal 2+ cases, and substantially lower at the 0/1+ boundary. The IHC unimodal baseline of 77.38% per source paper Table 1 is the immediate baseline to improve.

8. **Improve IHC unimodal branch accuracy using hybrid unfreezing strategy** (targeted at 20-percentage-point gap vs. literature ceiling per P001, P003/P004):

- **What to do**: Apply the hybrid unfreezing ResNet50 fine-tuning strategy from P001 (freeze Blocks 1–2 for generic histological features, fine-tune Blocks 3–4 and fc layer) to the MM'25 paper's IHC unimodal branch; implement the training configuration: Adam optimizer (lr=0.001), cosine annealing scheduler, early stopping (patience=5), L2 regularization (1×10^{-4}), data augmentation (horizontal flip, ±20° rotation, color jitter), batch size 32, 100 epochs max. - **Metric to measure**: IHC unimodal accuracy on BCI validation set; per-class F1 for each HER2 class; computational cost (GPU-hours per training run). - **Target to achieve**: IHC unimodal accuracy ≥ 90% (vs. current 77.38% per source paper Table 1), representing a ≥ 12.6 percentage point improvement; computational cost reduced by ≥ 70% vs. full fine-tuning (P001 documents 72.7% reduction). - **Baseline to compare against**: IHC unimodal baseline of 77.38% (Real IHC only, no fake H&E) per source paper Table 1. Literature ceiling of 97.9% for slide-level HER2 classification from IHC (per P003/P004 snippet). P001's ResNet50 hybrid unfreezing achieves 95% accuracy on BCI at 10×, with frozen backbone at 73% and full fine-tuning at 91% — use this as the architectural reference for the fine-tuning strategy.

8. Key Risks, Caveats, and Evidence Boundaries

1. **Single-institution evaluation risk** (named source: P010): All experiments use the BCI dataset from one institution. The 66-laboratory stain normalization study (P010) provides concrete evidence that cross-institution staining variation is substantial — histogram matching achieves intersection=0.891–0.944 across 66 labs, meaning 6–11% of color information is not aligned even with the best normalization method. The reported 94.25%

accuracy may not generalize to other institutions. Any clinical deployment decision must await external validation results.

2. **SSIM-accuracy dissociation mechanism uncertainty** (named source: P012): CM-GAN's cross-modal SSIM values (0.39–0.51) are substantially below SSIM > 0.90. The BCI pathologists' 37.5–40% accuracy interpreting generated IHC images (P012) confirms that synthesis artifacts are visually misleading to human experts. The dissociation between low SSIM and high classification accuracy may reflect artifact exploitation (H1) — in which case cross-institution deployment could see substantial accuracy drop as artifact patterns differ — or genuine shared-feature robustness (H2). Cross-institution evaluation is the only definitive discriminating experiment.

3. **Missing uncertainty quantification risk** (named source: P003, P004): The framework produces no confidence estimates. In clinical deployment, this means that incorrect predictions on borderline HER2 2+ or HER2 1+ cases cannot be identified and escalated for human review. P003 identifies the absence of uncertainty quantification as one of three major barriers to clinical deployment for HER2 AI systems. P007 demonstrates that calibration (ECE < 5%) is achievable in multimodal settings through logit-level mixture of experts.

4. **HER2-low boundary performance risk** (named source: P001, P003): The most therapeutically important category (HER2 1+, T-DXd eligible per P003) is not separately evaluated. P001 demonstrates F1 = 0.85 for HER2 1+ on BCI at 10×, substantially below aggregate accuracy. P003 reports inter-observer concordance of 60-75% for equivocal 2+ cases, indicating that even human experts find this boundary difficult. The MM'25 paper's aggregate accuracy may be inflated by strong performance on the easier HER2 0 and HER2 3+ classes.

5. **Evidence-transfer risk from non-histopathology literature** (named source: P006, P007, P008): Several key supporting papers (MoRA, SimMLM, MODES) are evaluated on non-histopathology domains (CXR, ODIR per P006; BraTS per P007; ECG/cMRI per P008). Direct transferability of their architectural principles to histopathology H&E+IHC image+image multimodal learning is unproven. MODES (P008) validates the shared-specific decoupling principle across domains, but the specific implementation (DCO/DAO losses, CBAM attention) may need domain-specific tuning for histopathology.

6. **Class imbalance risk** (named source: P004, P005): The paper mentions class-specific training weights but does not report the distribution of HER2 expression levels in the BCI dataset. P004 identifies limited training data diversity as one of three major barriers to HER2 AI clinical deployment. P005's ResViT-GANNet achieves 96.40% accuracy on multi-institutional BreakHis but shows 3.3% accuracy improvement specifically from GAN augmentation on

minority tumor classes, indicating that class imbalance materially affects performance.

7. **Generative model evaluation limitation** (named source: P010, P012): The CM-GAN's cross-modal reconstruction is evaluated only on the BCI single-institution dataset. P010 documents that deep learning-based stain normalization methods (CycleGAN, Pix2pix) produce hallucination artifacts (false nuclei in adipose tissue, miscolored smooth muscle nuclei) on H&E images. CM-GAN may produce analogous artifacts under cross-institution staining variation, but this relationship has not been analyzed. The pathologist accuracy of 37.5–40% on BCI-generated IHC (P012) confirms the artifact problem on single-institution data.

8. **No comparison with state-of-the-art HER2 AI systems** (named source: P001, P003, P004): The paper does not benchmark against established HER2 AI scoring tools. P001's hybrid unfreezing ResNet50 achieves 95% accuracy on BCI (vs. MM'25's 95.09% dual-modality), and P003/P004 report 97.9% accuracy for slide-level HER2 classification from IHC. The single-modality IHC baseline of 77.38% is substantially below the literature state-of-the-art, suggesting the IHC branch needs architectural improvement. The CM-GAN's 90.28% IHC reconstruction-assisted accuracy (vs. 77.38% unimodal) shows the value of cross-modal reconstruction, but the gap to the 97.9% ceiling indicates room for improvement.

Report generated from grounded analysis of ACM MM'25 paper (arXiv:2506.10006v2) and 12 supporting literature papers.