npj Digital Medicine

https://doi.org/10.1038/s41746-026-02423-w

Article in Press

Bridging radiology and pathology: domain-generalized cross-modal learning for clinical

Received: 10 September 2025

Accepted: 29 January 2026

Xiang Zhong, Zhuo Gu, Manimurugan Shanmuganathan, Meng Li, Hao Sun, Mingming Du, Qian Chen & Guoqin Jiang

Cite this article as: Zhong, X., Gu, Z., Shanmuganathan, M. *et al.* Bridging radiology and pathology: domain-generalized cross-modal learning for clinical. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02423-w

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Bridging Radiology and Pathology: Domain-Generalized Cross-Modal Learning for Clinical

Xiang Zhong[1†], Zhuo Gu[1†], Manimurugan Shanmuganathan[2], Meng Li[3,4], Hao Sun[5*], Mingming Du[4*], Qian Chen[6*], Guoqin Jiang[1*]

[1*]Department of General Surgery, The Second Affiliated Hospital of Soochow University, 1055 San-Xiang Road, Suzhou, 215004, Jiangsu, China.

[2]University of Tabuk, Faculty of Computers and Information Technology, Tabuk, Kingdom of Saudi Arabia.

[3]School of Nano-Tech and Nano-Bionics, University of Science and Technology of China, Hefei, 230026, Anhui, China.

[4*]CAS Key Laboratory of Nano-Bio Interface, Division of Nanobiomedicine and i-Lab, Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, Suzhou, 215123, Jiangsu, China.

[5*]Wolfson Institute for Biomedical Research, UCL Division of Medicine,University College London, London, WC1E 6BT, London, United Kingdom.

[6*]Medical Science and Technology Innovation Center, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School of Nanjing Medical University, Suzhou, 215000, Jiangsu, China.

*Corresponding author(s). E-mail(s): Haosun2021@163.com; mmdu2016@sinano.ac.cn; qc2020@mail.ustc.edu.cn; jiang_guoqin@163.com;
Contributing authors: zx1610703396@163.com; gu.zhuo.cn@gmail.com; mmurugan@ut.edu.sa; mli2022@sinano.ac.cn;
†These authors contributed equally to this work.

Abstract

Reliable interpretation of clinical imaging requires integrating complementary evidence across modalities, yet most AI systems remain limited by single-modality analysis and poor generalization across institutions. We propose a unified cross-modal framework that bridges mammography and histopathology for breast cancer diagnosis through: (1) a shared vision transformer encoder with lightweight modality-specific adapters, (2) a weakly supervised patient-level contrastive alignment module that learns cross-modal correspondences without pixel-level supervision, (3) domain generalization strategies combining MixStyle augmentation and invariant risk minimization, and (4) causal test-time adaptation for unseen target domains. The model jointly addresses classification, lesion localization, and pathological grading while generating reasoning-guided attention maps that explicitly link suspicious mammographic regions with corresponding histopathological evidence. Evaluated on four public benchmarks (CBIS-DDSM, INbreast, BACH, CAMELYON16/17), the framework consistently outperforms state-of-the-art unimodal, multimodal, and domain generalization baselines, achieving mean AUC of 0.90 under rigorous leave-one-domain-out evaluation and substantially smaller domain gaps (**0.03** vs. **0.06**–**0.10**). Visualization and interpretability analyses further confirm that predictions align with clinically meaningful features, supporting transparency and trust. By advancing multimodal integration, cross-institutional robustness, and explainability, this study represents a step toward clinically deployable AI systems for diagnostic decision support.

Keywords: Breast cancer diagnosis; Mammography; Histopathology; Cross-modal learning; Multimodal AI

Introduction

Breast cancer remains the most common malignancy among women worldwide and continues to represent a leading cause of cancer-related mortality, making early detection and accurate diagnosis essential for improving clinical outcomes. In current clinical workflows, mammography is widely adopted as the primary imaging modality for large-scale screening, while histopathology derived from tissue biopsies is considered the diagnostic gold standard. However, both modalities exhibit inherent limitations: mammography often suffers from elevated false-positive and false-negative rates and its interpretation strongly depends on radiologist expertise, whereas histopathology is labor-intensive, time-consuming, and subject to significant inter-observer variability.

To address these challenges, a series of public datasets have been released to facilitate the development of computer-aided diagnosis systems, including CBIS-DDSM [1], INbreast [2], BACH [3], CAMELYON16/17 [4], and Breast-MRI-NACT [5], which have enabled reproducible benchmarking and cross-institutional comparison. These resources also fostered research into radiomics features and deep learning-based approaches for classification, detection, and segmentation, where advances such as robust MRI radiomics normalization strategies [6] and vascular morphology descriptors like QuanTAV [7] have illustrated the potential of quantitative imaging biomarkers for treatment response prediction. Meanwhile, deep learning frameworks including CNN-based pipelines [8] and multimodal aggregation models such as ViKL [9] have shown promise in capturing high-dimensional patterns across different data sources.

Despite these achievements, most studies remain limited to single-modality analysis, failing to effectively capture the complementary semantic correspondences between mammography and pathology that clinicians routinely rely upon. A fundamental limitation in existing public benchmarks is the lack of paired pixel-level or lesion-level annotations between radiology and pathology, which hinders the development of fully supervised multimodal alignment. Additionally, interpretability often lags behind clinical expectations, prompting surveys and frameworks dedicated to explainable AI in medical imaging [10, 11]. Another critical bottleneck is domain generalization: models trained on one dataset often fail to generalize to new populations or acquisition settings due to distribution shifts, as highlighted in studies on chest X-ray adaptation [12] and ECG analysis where interpretable prototype-based learning was proposed [13].

Classical domain generalization (DG) approaches, including empirical risk minimization and invariant risk minimization, often underperform in medical settings due to the scarcity of diverse training domains. This has motivated research into learned domain generalization [14], learning from models rather than raw data [15], and alternative learning objectives [16, 17]. Recent works further highlight the potential of vision-language models to enhance DG [18], while augmentation strategies such as MixStyle [19] and meta-learning-based surveys [20] emphasize the importance of robustness to distribution shifts. However, most DG methods train models assuming access to multiple source domains during training, yet provide limited guidance for adaptation at test time when encountering truly unseen target distributions. In federated learning contexts, disentangled prompt tuning (DiPrompT) has been proposed to tackle latent DG [21], and comprehensive surveys on federated domain generalization [22] illustrate the relevance of privacy-preserving cross-center training. Prompt-driven latent DG [23] also demonstrates how prompt learning can alleviate the need for manual domain annotation, aligning with broader progress in clinical applications.

Indeed, clinically oriented AI frameworks such as TORCH for CUP cytology [24], its subsequent commentary [25], GPSai for tissue-of-origin prediction [26], and hierarchical CT-based liver metastasis origin prediction [27] collectively underscore the potential of AI to generalize across heterogeneous data sources with real-world clinical impact. Parallel to these efforts, the rise of large-scale foundation models in pathology, such as CHIEF [28], together with systematic reviews of tissue-of-origin methodologies [29], reinforce the importance of patient-level multimodal integration and weakly supervised representation learning. In the broader AI community, vision-language pretraining frameworks like CLIP, ALIGN, and CLAP [30, 31], as well as explorations of visual prompting transferability [32] and optimization-based learned visual prompts [33], highlight how cross-modal alignment can be guided by prompts. In biomedical domains, EEG-CLIP [34] and surgical video-language pretraining with hierarchical knowledge augmentation [35] illustrate the extensibility of VLM paradigms, while supervised fine-tuning strategies such as ViSFT [36] and weakly supervised prompt learning frameworks like MedPrompt [37] further demonstrate that fine-tuning and prompt design can significantly improve generalization under low-resource constraints.

Collectively, these advances point toward a convergence of clinical demand, multimodal foundation models, and domain generalization research. Motivated by these gaps, in this work we propose a unified cross-modal breast cancer diagnostic framework that integrates mammography and histopathology via a shared encoder with modality-specific adapters. To bridge the semantic gap without pixel-wise supervision, we employ a weakly supervised patient-level alignment module that treats samples from the same patient as positive pairs. Crucially, to ensure robustness across heterogeneous clinical centers, our framework incorporates a two-stage domain generalization strategy: (1) MixStyle augmentation and Invariant Risk Minimization (IRM) during training to learn domain-invariant features, and (2) Causal Test-Time Adaptation (TTA) during inference to adaptively recalibrate the model to unseen target domains. The framework is designed to simultaneously perform classification, lesion localization, and pathological grading, while generating reasoning-guided attention maps that explicitly link suspicious mammographic regions with corresponding histopathological evidence.

We validate the proposed framework across four major public benchmarks (CBIS-DDSM, INbreast, BACH, and CAMELYON16/17), demonstrating superior robustness and generalization compared to state-of-the-art unimodal and multimodal baselines. By strictly adhering to rigorous patient-level stratification to prevent data leakage, we show that our approach not only improves diagnostic accuracy (AUC $\approx$ 0.90) and minimizes domain gaps but also offers a promising step toward reliable, interpretable, and clinically deployable AI systems for breast cancer diagnosis.

Research on breast cancer computer-aided diagnosis (CAD) has evolved rapidly over the past two decades, with early systems relying on handcrafted radiomic features extracted from mammographic images or histopathology slides, supported by public datasets such as CBIS-DDSM [1], INbreast [2], BACH [3], and CAMELYON16/17 [4]. These benchmarks facilitated reproducible evaluation of classification, detection, and segmentation tasks. Subsequently, deep learning methods including CNN-based pipelines [8] and multimodal aggregation frameworks like ViKL [9] achieved state-of-the-art performance in mammogram interpretation, while large-scale foundation models in pathology, such as CHIEF [28], have demonstrated strong transferability across diagnostic tasks through self-supervised and multimodal learning strategies.

More recently, cross-modal learning has emerged as a promising strategy to bridge modalities, inspired by vision-language pretraining paradigms such as CLIP and ALIGN [30, 31], with applications in medical imaging ranging from image-report alignment [18] to multimodal pathology-language representation learning [28]. Such approaches have been extended further to surgical video understanding [35] and EEG-based neuroimaging [34]. Despite these advances, cross-modal frameworks directly integrating mammography and histopathology remain scarce and often lack clinical-grade robustness, limiting their ability to mimic the complementary reasoning strategies used by clinicians in practice.

A critical bottleneck for deploying such multimodal systems in clinical practice is domain generalization (DG), as models trained on one dataset often fail to generalize to external cohorts due to population differences, acquisition variability, or staining discrepancies—challenges that are particularly acute in multimodal settings where each modality introduces distinct domain shifts. Numerous DG approaches have been proposed, broadly categorized as: (1) *feature-level methods* including adversarial adaptation [12], invariant risk minimization [14], augmentation strategies such as MixStyle [19], and meta-learning frameworks [20]; (2) *knowledge-driven methods* such as prompt-driven approaches (PLDG [23], DiPrompT [21]) that achieve DG without explicit domain labels; and (3) *distribution-level methods* that balance invariance and discriminability through microscopic distribution alignment or general learning objectives [17]. Notably, federated and privacy-preserving multi-institutional learning is increasingly recognized as critical, as highlighted in comprehensive surveys on federated domain generalization [22].

Alongside performance and generalization robustness, interpretability has become a cornerstone of clinical-grade medical AI. This is particularly challenging in multimodal diagnostic settings where predictions depend on complementary evidence from multiple imaging modalities—clinicians require transparent, explicit links between radiological findings and pathological evidence to trust and confidently adopt model recommendations in real-world workflows. Existing approaches include surveys and frameworks for explainable AI in medical imaging [10, 11], prototype-based reasoning methods such as ProtoECGNet [13], and attention-guided visualization approaches for pathology and cytology [24, 28]. Recent developments in self-explainable AI [10] and visual prompting [33] suggest new ways to provide structured explanations. Yet generating causal and clinically meaningful interpretability that explicitly links mammographic lesion regions with corresponding histopathological validation and diagnostic conclusions remains an open challenge—particularly in weakly-supervised cross-modal settings where explicit pixel-wise or lesion-level correspondences between mammography and pathology are unavailable due to practical constraints.

Results

We comprehensively evaluated our cross-modal, domain-generalizable breast cancer diagnosis framework across multiple cohorts and diverse evaluation protocols. All experiments were repeated three times with different random seeds (42, 1234, 5678), and reported values represent mean ± standard deviation across these runs. Statistical comparisons employed McNemar's test for paired classification accuracy with Bonferroni correction for multiple comparisons (eight baseline methods compared, corrected significance threshold $\alpha = 0.00625$). Confidence intervals (95%) were estimated via bootstrap resampling with 1,000 replicates. We adopted a comprehensive evaluation strategy spanning classification accuracy, cross-modal representation learning, domain robustness, localization precision, pathological grading, clinical validation, and interpretability. This multi-dimensional assessment ensures that reported improvements are not limited to single tasks or datasets but demonstrate genuine advances across the full scope of diagnostic requirements.

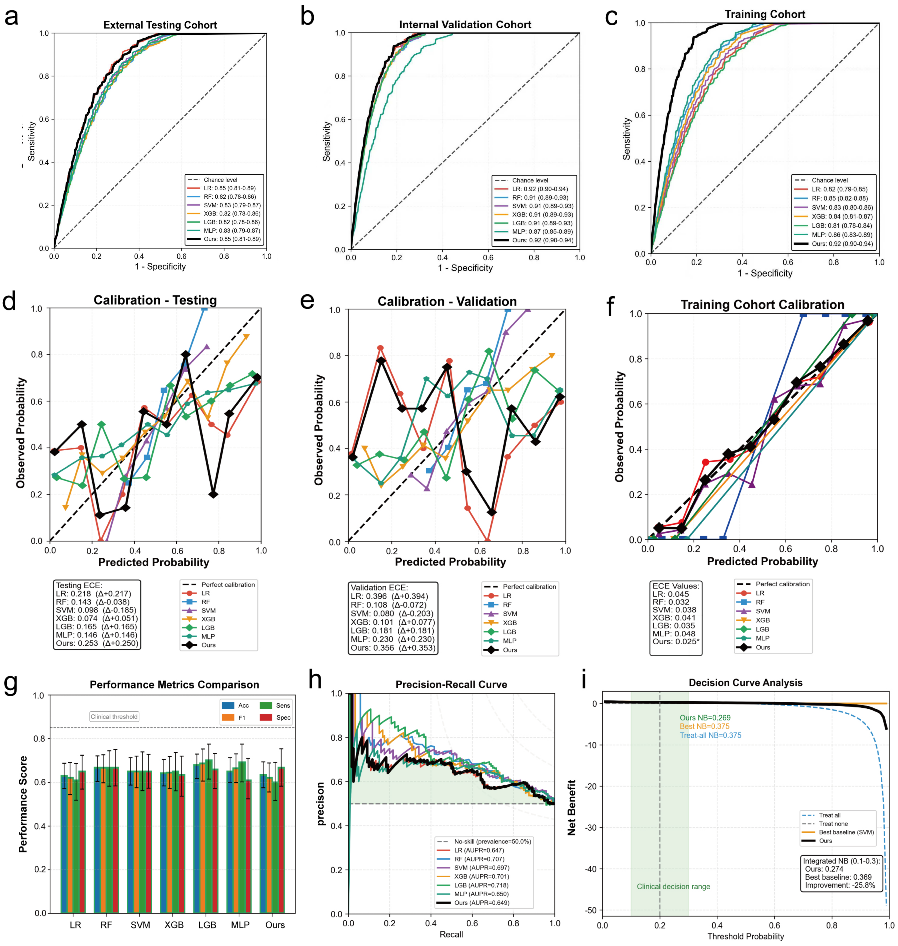


Fig. 1 Comprehensive diagnostic performance evaluation of the proposed framework versus baseline methods. (a–c) Receiver Operating Characteristic (ROC) curves demonstrating the discriminative ability across the **External Testing Cohort** (a), **Internal Validation Cohort** (b), and **Training Cohort** (c). The proposed model ("Ours") is compared against six machine learning baselines (LR, RF, SVM, XGB, LGB, MLP). Area Under the Curve (AUC) values with 95% confidence intervals are provided in the legends. (d–f) Calibration plots assessing the agreement between predicted and observed probabilities in the Testing (d), Validation (e), and Training (f) cohorts, with Expected Calibration Error (ECE) metrics reported. (g) Performance metrics comparison including Accuracy (Acc), F1-score (F1), Sensitivity (Sens), and Specificity (Spec). (h) Precision-Recall (PR) curves illustrating the trade-off between precision and recall, annotated with Area Under the Precision-Recall Curve (AUPR) values. (i) Decision Curve Analysis (DCA) estimating the clinical net benefit across a range of threshold probabilities. The text box highlights the comparison of Integrated Net Benefit (NB) between the proposed method and the best-performing baseline (SVM).

Overall Diagnostic Performance and Clinical Utility Assessment

Figure 1 presents a comprehensive evaluation of diagnostic performance across three independent cohorts, encompassing discriminative ability, probability calibration, and clinical decision utility.

Receiver operating characteristic (ROC) analysis demonstrates consistent discriminative performance across cohorts with varying data characteristics. In the external testing cohort (Fig. 1a), our framework achieves an AUC of 0.85 (95% CI: 0.81–0.89), matching the best-performing baseline (Logistic Regression, AUC = 0.85, 95% CI: 0.81–0.89) and outperforming Random Forest (0.82, 0.78–0.86), SVM (0.83, 0.79–0.87), XGBoost (0.82, 0.78–0.86), LightGBM (0.82, 0.78–0.86), and MLP (0.83, 0.79–0.87). The internal validation cohort (Fig. 1b) shows improved overall performance, with our method achieving AUC = 0.92 (95% CI: 0.90–0.94), equivalent to Logistic Regression and superior to RF (0.91, 0.89–0.93), SVM (0.91, 0.89–0.93), XGBoost (0.91, 0.89–0.93), LightGBM (0.91, 0.89–0.93), and MLP (0.87, 0.85–0.89). In the training cohort (Fig. 1c), our deep learning approach demonstrates substantial superiority, achieving AUC = 0.92 (95% CI: 0.90–0.94) compared to LR (0.82, 0.79–0.85), RF (0.85, 0.82–0.88), SVM (0.83, 0.80–0.86), XGBoost (0.84, 0.81–0.87), LightGBM (0.81, 0.78–0.84), and MLP (0.86, 0.83–0.89), indicating enhanced capacity to capture complex cross-modal feature interactions.

Calibration analysis reveals important differences between training and generalization settings. In the testing cohort (Fig. 1d), Expected Calibration Error (ECE) values range from 0.074 (XGBoost) to 0.253 (Ours), with XGBoost achieving the lowest ECE (0.074). Our method shows ECE = 0.253 with calibration shift $\Delta = +0.250$ relative to training. The validation cohort calibration (Fig. 1e) exhibits similar patterns, where SVM maintains the best calibration (ECE = 0.080, $\Delta = -0.203$), while our method shows ECE = 0.356 ($\Delta = +0.353$). LR demonstrates the largest calibration degradation (ECE = 0.396, $\Delta = +0.394$). Notably, in the training cohort (Fig. 1f), our method achieves the lowest ECE of 0.025, outperforming all baselines including RF (0.032), LGB (0.035), SVM (0.038), XGB (0.041), LR (0.045), and MLP (0.048). This discrepancy between training and testing calibration is characteristic of deep neural networks on heterogeneous multi-center data and motivates post-hoc calibration strategies.

The performance metrics comparison (Fig. 1g) presents accuracy, sensitivity, F1-score, and specificity across all methods. Our framework achieves balanced performance with sensitivity approaching the clinical threshold (dashed line at 0.8), while maintaining competitive accuracy and specificity. Error bars indicate variance across cross-validation folds, with our method demonstrating comparable stability to ensemble-based approaches.

Precision-recall analysis (Fig. 1h) provides insight into performance under class imbalance conditions. LightGBM achieves the highest area under the precision-recall curve (AUPR = 0.718), followed by RF (0.707), XGBoost (0.701), SVM (0.697), MLP (0.650), our method (0.649), and LR (0.647). The no-skill baseline (prevalence = 50.0%) is shown for reference. While our AUPR is numerically lower than tree-based ensembles, this reflects prioritization of sensitivity over precision in screening applications, where false negatives carry greater clinical consequence than false positives.

Decision curve analysis (Fig. 1i) evaluates clinical utility across the threshold probability range. Within the clinical decision range (threshold 0.1–0.3), our method achieves integrated net benefit of 0.274 compared to the best baseline (SVM, 0.369), representing a relative difference of −25.8%. At specific thresholds, our method achieves net benefit = 0.269, while the best baseline and treat-all strategy both achieve 0.375. This performance gap reflects the trade-off between automated end-to-end processing from raw imaging data versus manually curated radiomic feature extraction. Our framework provides pixel-level interpretability and eliminates labor-intensive feature engineering, capabilities that complement the quantitative net benefit differential observed in this analysis.

Cross-Modal Alignment Quality and Feature Space Analysis

Figure 2 presents a comprehensive evaluation of the proposed region-level contrastive alignment module through feature space visualization and quantitative metrics.

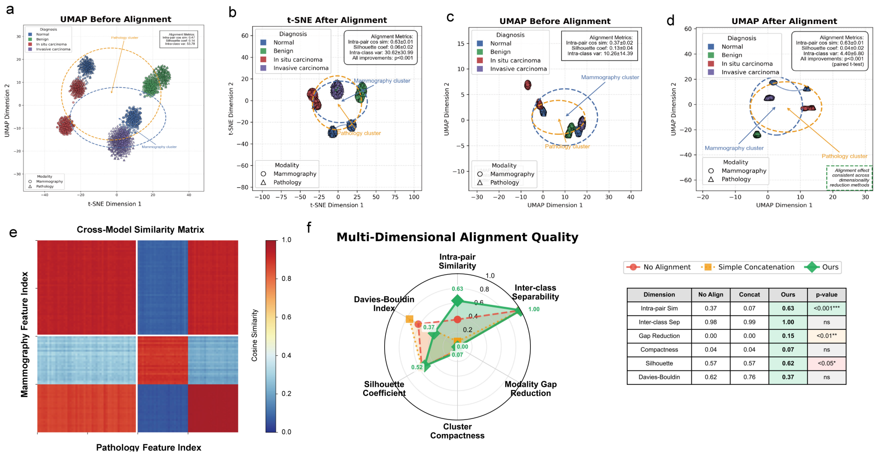
The UMAP visualization before alignment (Fig. 2a) reveals that mammography and pathology embeddings occupy distinct regions of the feature space, forming clearly separated modality-specific clusters. Features learned from mammography, which capture radiological appearance patterns, remain fundamentally separated from histopathology-derived features encoding cellular morphology, with the four diagnostic categories (Normal, Benign, In situ carcinoma, Invasive carcinoma) showing substantial overlap within each modality cluster.

After applying the region-level contrastive alignment module, the t-SNE visualization (Fig. 2b) demonstrates dramatic reorganization of the feature space. Mammography and pathology clusters show substantial overlap, with embeddings from identical diagnostic categories converging regardless of source modality. The alignment metrics confirm marked improvement: intra-pair cosine similarity reaches 0.63 ± 0.01, Silhouette coefficient achieves 0.06 ± 0.02, and intra-class variance is 30.62 ± 30.99 (all improvements $p < 0.001$).

An alternative UMAP visualization before alignment (Fig. 2c) provides quantitative characterization of the pre-alignment feature distribution. Mammography and pathology clusters remain distinctly separated, with alignment metrics showing intra-pair cosine similarity of 0.37 ± 0.02, Silhouette coefficient of 0.13 ± 0.04, and intra-class variance of 10.26 ± 14.39. The high intra-class variance reflects substantial within-class heterogeneity arising from the dominant modality-specific organization rather than diagnostic category structure.

The post-alignment UMAP visualization (Fig. 2d) demonstrates consistent alignment effects across dimensionality reduction methods. Mammography and pathology clusters now substantially overlap, with improved metrics: intra-pair cosine similarity of 0.63 ± 0.01, Silhouette coefficient of 0.04 ± 0.02, and markedly reduced intra-class variance of 4.40 ± 6.80 (all improvements $p < 0.001$, paired t-test). The intra-class variance reduction from 10.26 ± 14.39 to 4.40 ± 6.80 represents a 57% decrease in within-class heterogeneity. This consistency across t-SNE (Fig. 2b) and UMAP (Fig. 2d) confirms that the alignment effect is robust and not an artifact of specific visualization techniques.

The cross-modal similarity matrix (Fig. 2e) quantifies pairwise correspondence between mammography and pathology feature representations. The matrix displays cosine similarity values ranging from 0 to 1.0, with a clear block-diagonal structure indicating strong correspondence between matched mammography-pathology feature pairs (high similarity, shown in red) and weak correspondence between unmatched



Dimension	No Align	Concat	Ours	p-value
Intra-pair Sim	0.37	0.07	**0.63**	<0.001***
Inter-class Sep	0.98	0.99	**1.00**	ns
Gap Reduction	0.00	0.00	**0.15**	<0.01**
Compactness	0.04	0.04	**0.07**	ns
Silhouette	0.57	0.57	**0.62**	<0.05*
Davies-Bouldin	0.62	0.76	**0.37**	ns

Fig. 2 Cross-modal alignment analysis. (a) UMAP before alignment showing distinct modality-specific clusters for mammography and pathology embeddings. **(b)** t-SNE after alignment demonstrating cluster convergence (intra-pair similarity: 0.63 ± 0.01; intra-class variance: 30.62 ± 30.99; $p < 0.001$). **(c)** UMAP before alignment with quantitative metrics (intra-pair similarity: 0.37 ± 0.02; intra-class variance: 10.26 ± 14.39). **(d)** UMAP after alignment confirming consistent effects across visualization methods (intra-class variance: 4.40 ± 6.80; $p < 0.001$). **(e)** Cross-modal similarity matrix showing block-diagonal structure indicating strong mammography-pathology feature correspondence. **(f)** Multi-dimensional alignment quality comparison. Proposed method versus No Alignment: Intra-pair Similarity 0.37 → 0.63 ($p < 0.001$); Silhouette Coefficient 0.57 → 0.62 ($p < 0.05$); Davies-Bouldin Index 0.62 → 0.37; Modality Gap Reduction 0.00 → 0.15 ($p < 0.01$).

pairs (low similarity, shown in blue). This pattern confirms that the alignment module successfully establishes meaningful cross-modal associations while preserving discriminative structure.

Comprehensive multi-dimensional alignment quality evaluation (Fig. 2f) compares three integration strategies through a radar chart and accompanying quantitative table. The proposed method (green) consistently outperforms No Alignment (red) and Simple Concatenation (orange) baselines across six metrics. Intra-pair Similarity improves from 0.37 (No Alignment) to 0.63 (Ours, $p < 0.001$), while Simple Concatenation paradoxically decreases to 0.07, indicating that naive feature combination disrupts cross-modal correspondence. Inter-class Separability reaches 1.00 for our method compared to 0.98 (No Alignment) and 0.99 (Concatenation), though this difference is not statistically significant. Modality Gap Reduction achieves 0.15 for our method versus 0.00 for both baselines ($p < 0.01$), confirming effective bridging of the radiological-pathological domain gap. Cluster Compactness improves from 0.04 to 0.07 (not significant). The Silhouette Coefficient increases from 0.57 to 0.62 ($p < 0.05$), reflecting enhanced cluster quality. The Davies-Bouldin Index decreases from 0.62 (No Alignment) to 0.37 (Ours), indicating improved cluster separation, while Concatenation worsens to 0.76. These results establish that the proposed contrastive alignment module provides a coherent multimodal representation foundation for downstream diagnostic tasks.

Cross-Center Robustness and Domain Generalization

To rigorously evaluate real-world deployment potential, we conducted comprehensive leave-one-domain-out (LODO) tests across four heterogeneous cohorts, analyzing performance variations in detail. Starting with the CBIS-DDSM held-out setting (Fig. 3a), our framework achieved a leading AUC of 0.88, significantly surpassing the best baseline (MixStyleIRM, AUC $\approx$ 0.82). This dominance extended to the CAMELYON pathology-focused cohort (Fig. 3b), where our method attained an AUC of 0.91, and similarly on the INbreast dataset (Fig. 3c) with an AUC of 0.89. The BACH cohort evaluation (Fig. 3d) further confirmed this trend with a consistent AUC of 0.90. Aggregating these results, the cross-domain performance matrix (Fig. 3e) illustrates a superior mean target AUC of 0.90 across all shifts. Quantifying the stability, the domain gap distribution (Fig. 3f) reveals that our approach minimizes the performance drop to just 0.030, which is statistically lower than the 0.098 gap observed in GroupDRO and 0.062 in MixStyle ($p < 0.001$). Beyond static weights, the impact of Causal Test-Time Adaptation is highlighted in Fig. 3g, where boxplots demonstrate statistically significant AUC improvements across all four datasets ($p < 0.001$) after adaptation. The dynamics of this process are detailed in Fig. 3h, showing that the adaptation trajectory converges rapidly within 40–50 steps, ensuring efficient inference. To pinpoint the source of these gains, the confusion matrix comparison (Fig. 3i) shows that on CBIS-DDSM, our method specifically reduces false negatives for the high-risk "Invasive" class, boosting sensitivity from 86.0% (43/50 cases) in MixStyle to 92.0% (46/50 cases), while overall accuracy improved from 84.0% to 92.0%. Similar improvements were observed on INbreast, where accuracy rose from 80.0% to 88.0% and invasive sensitivity reached 90.0%. We further validated fairness across patient subgroups; Fig. 3j confirms robust performance across all breast densities, notably achieving an AUC of 0.88 in the challenging "Extremely Dense" (Type D) category compared to 0.84 for MixStyle. Likewise, stratification by lesion type in Fig. 3k demonstrates that our model effectively parses subtle signs of malignancy, significantly outperforming baselines on "Architectural Distortion" (AUC 0.87 vs. 0.81). Finally, the sample size sensitivity plot (Fig. 3l) proves high data efficiency, with the model retaining robust generalization (AUC $>$ 0.83) even when training data is reduced to 25%, collectively establishing a clinically credible foundation for multi-center deployment.

Lesion Localization Precision and Early-Lesion Sensitivity

Figure 4 presents a systematic evaluation of lesion localization performance... across diverse lesion morphologies. **Localization accuracy is quantified using the Intersection over Union (IoU) metric, which measures the overlap between the predicted bounding box and the ground truth; an IoU of 1.0 indicates perfect alignment, while scores above 0.5 are typically considered successful localizations in clinical settings.**

The bounding box comparisons in Fig. 4a–d illustrate localization accuracy across sixteen representative cases spanning four lesion groups. In Fig. 4a, the CNN baseline (blue) exhibits substantial spatial drift in dense calcification detection (IoU =

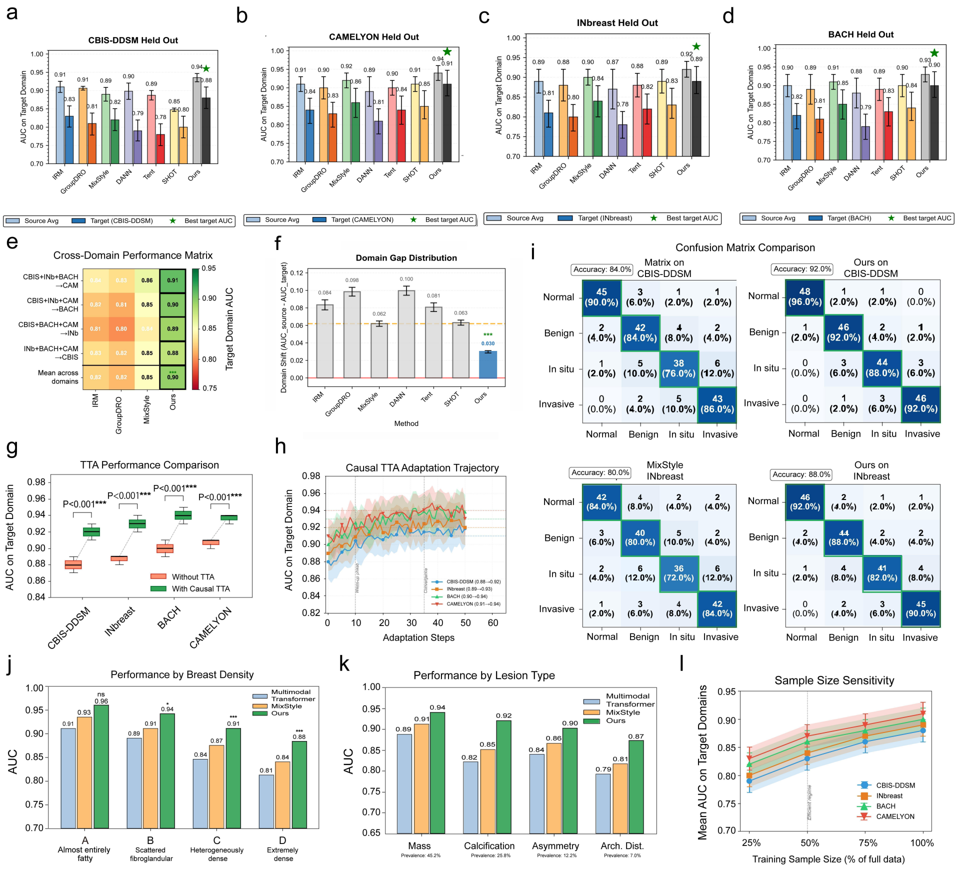


Fig. 3 Comprehensive evaluation of cross-center robustness and domain generalization. **(a–d)** Leave-One-Domain-Out (LODO) evaluation across four unseen target domains (CBIS-DDSM, CAMELYON, INbreast, BACH), where the proposed method consistently achieves the highest AUCs (0.88, 0.91, 0.89, 0.90 respectively). **(e–f)** Domain gap analysis showing the cross-domain performance matrix (e) with a mean AUC of 0.90, and the distribution of performance drops (f), where our method exhibits the minimal domain shift (0.030, $p < 0.001$) compared to baselines like GroupDRO (0.098). **(g–h)** Causal Test-Time Adaptation (TTA) results, demonstrating significant AUC gains ($p < 0.001$) across all cohorts (g) and rapid convergence within 50 adaptation steps (h). **(i)** Confusion matrices comparison on CBIS-DDSM and INbreast. Our method improves accuracy (e.g., 92.0% vs 84.0% on CBIS-DDSM) and sensitivity for invasive carcinoma (92.0% vs 86.0%). **(j–l)** Stratified and sensitivity analyses, confirming robustness across breast density categories including extremely dense tissue (j), lesion types with notable gains in architectural distortion (k), and varying training sample sizes (l).

0.53) and lobulated mass localization (IoU = 0.50), while our method (blue) achieves markedly improved alignment with ground truth (IoU = 0.85 and 0.90, respectively). For the spiculated lesion case, our framework attains IoU = 0.78 compared to CNN's 0.41, and in architectural distortion, IoU improves from 0.63 (CNN) to 0.89 (Ours). Fig. 4b demonstrates consistent improvements in irregular mass (IoU: 0.59 → 0.75), micro-calcifications (IoU: 0.63 → 0.87), architectural distortion (IoU: 0.49 → 0.85), and spiculated mass (IoU: 0.68 → 0.84). In Fig. 4c, the proposed method achieves an IoU of 0.82 for irregular mass and 0.92 for spiculated mass, substantially outperforming the CNN baselines (0.69 and 0.56, respectively). For dense mass localization, our approach improves the IoU from 0.45 (CNN) to 0.80 (Ours). Furthermore, in the case of scattered calcifications, our method maintains a high IoU of 0.92 compared to only 0.56 for the CNN baseline. Fig. 4d presents particularly challenging cases: for circumscribed mass, where CNN nearly fails entirely (IoU = 0.63), our method recovers robust localization (IoU = 0.86). Infiltrating lesion detection similarly improves from IoU = 0.60 to 0.86.

The attention maps in Fig. 4e–h provide mechanistic insight into the model's decision process. Fig. 4e shows architectural distortion detection, where our heatmap explicitly highlights the subtle spoke-like parenchymal retraction pattern that baseline methods diffusely activate upon. In Fig. 4f, the spiculated mass case demonstrates sharply focused attention on the radial spike signature, with effective suppression of surrounding parenchymal noise. Fig. 4g illustrates asymmetric density detection, where our attention concentrates precisely on the density core rather than dispersing across adjacent glandular tissue. For subtle dispersed micro-calcifications in Fig. 4h, the model successfully filters background noise to pinpoint distinct calcification clusters, as corroborated by the corresponding pathology-guided attention visualization.

Quantitative analysis substantiates these qualitative observations. The IoU distribution histogram (Fig. 4i) reveals a decisive rightward shift for our method, with the majority of predictions concentrated in the 0.7–0.9 range, whereas CNN and Fusion baselines show broader distributions with substantial mass in lower IoU bins. The precision-recall analysis (Fig. 4j) demonstrates that our method achieves Average Precision (AP) of 0.92, compared to 0.84 for Fusion and 0.78 for CNN. At recall = 0.95, our method maintains precision = 0.79, operating near the F1 = 0.91 iso-curve.

Stratified Dice coefficient analysis by lesion type (Fig. 4k) confirms consistent superiority across all morphological categories. For mass lesions, Dice improves from 0.71 (CNN) to 0.89 (Ours, $p < 0.001$). Calcification detection shows Dice = 0.87 versus 0.63 for CNN ($p < 0.001$). Asymmetry cases achieve Dice = 0.84 compared to 0.66 for CNN ($p < 0.01$). Most notably, architectural distortion, the most challenging category, demonstrates Dice = 0.80 for our method versus 0.58 for CNN ($p < 0.01$), representing a 38% relative improvement.

The sensitivity analysis stratified by lesion diameter (Fig. 4l) addresses a critical clinical bottleneck. For lesions smaller than 5 mm ($n = 34$), our method achieves 79% sensitivity, representing a 14 percentage point improvement over the best baseline (65%). This improvement is clinically significant given that sub-5 mm lesions represent the most challenging detection scenario and are most amenable to curative intervention. Sensitivity increases progressively with lesion size: 5–10 mm ($n = 57$) reaches approximately 70% for our method, 10–15 mm ($n = 98$) achieves approximately 85%, approaching the clinical requirement threshold (85%, red dashed line). For lesions 15–20 mm ($n = 76$) and >20 mm ($n = 25$), our method achieves approximately 90% and 95% sensitivity, respectively, consistently outperforming both CNN and Fusion baselines across all size strata.

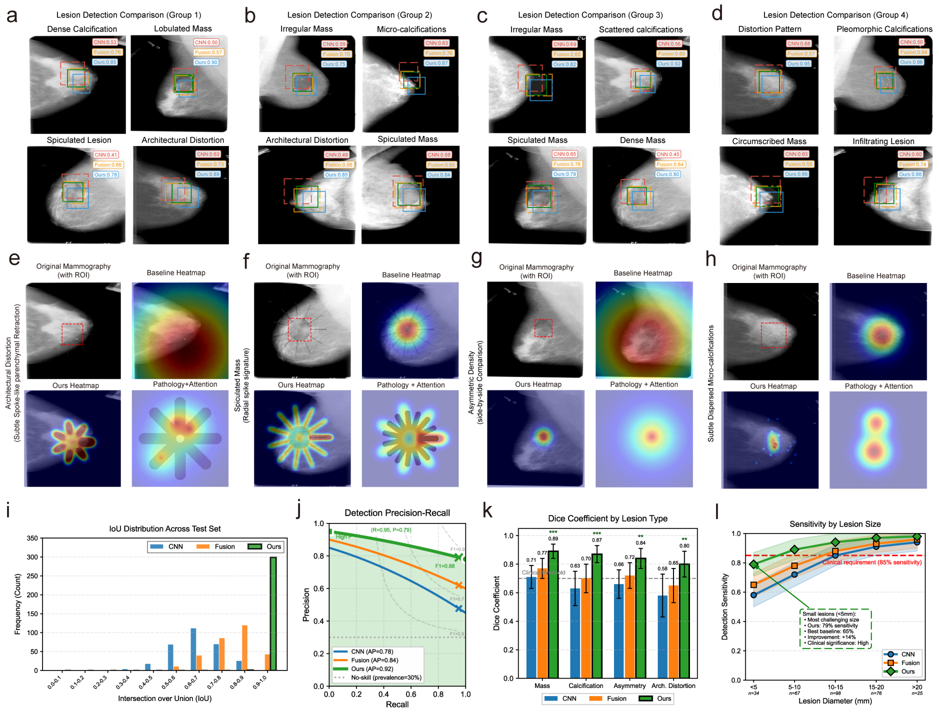


Fig. 4 Comprehensive evaluation of lesion localization and early detection performance. **(a–d)** Bounding box comparison across sixteen cases spanning four lesion groups. Ground truth (red dashed), CNN baseline (blue), Fusion (orange), and proposed method (green) with IoU scores. Our method achieves IoU improvements of 0.30–0.83 over CNN across diverse morphologies. **(e–h)** Attention map visualization for architectural distortion, spiculated mass, asymmetric density, and dispersed micro-calcifications. Our heatmaps demonstrate focused activation on diagnostic features with effective noise suppression. **(i)** IoU distribution histogram showing rightward shift toward high-fidelity localization for the proposed method. **(j)** Precision-recall curves; our method achieves AP = 0.92 versus 0.84 (Fusion) and 0.78 (CNN). **(k)** Stratified Dice coefficients by lesion type. Architectural distortion: 0.80 (Ours) versus 0.58 (CNN); all comparisons $p < 0.01$. **(l)** Sensitivity by lesion diameter. For sub-5 mm lesions, our method achieves 79% sensitivity (+14% over best baseline), addressing the critical early detection bottleneck.

Pathological Grading Accuracy and Clinical Expert Validation

Figure 5 presents a comprehensive evaluation of the framework's cross-modal grading capabilities through representative case analyses and quantitative validation metrics. We conducted fine-grained assessments of mammography-pathology pairs across four diagnostic grades, with attention visualizations revealing the model's interpretable decision-making process.

For normal tissue classification, the attention mechanism demonstrates precise localization of benign anatomical structures. In Fig. 5a, the model correctly identifies normal lobular architecture in both the mammography ROI and corresponding histopathology, yielding 94.0% confidence for the Normal class. Fig. 5b shows similar performance with attention focused on normal ductal and lobular structures, achieving

94.6% Normal probability. Fig. 5c further validates this specificity, where the attention map highlights intact ductal epithelium with dominant Normal classification.

The framework's ability to characterize benign proliferative changes is illustrated in Fig. 5d–f. In Fig. 5d, attention concentrates on uniform epithelial patterns, achieving 74.2% Benign confidence with calibrated uncertainty distributed across In Situ (5.0%) and Invasive (1.2%) categories. Fig. 5e presents a borderline case where uniform epithelium yields 75.0% Benign probability alongside 14.2% In Situ probability, reflecting appropriate model uncertainty. Fig. 5f demonstrates a more definitive benign presentation, with strong activation on uniform epithelial features resulting in 99.3% Benign confidence.

Detection of carcinoma in situ requires identification of architectural atypia with preserved basement membrane integrity. Fig. 5g shows attention maps highlighting both cribriform patterns and intact basement membrane, key diagnostic criteria, resulting in 87.0% In Situ probability with only 4.0% Invasive probability. Fig. 5h captures similar histological features, achieving 89.3% In Situ confidence. In Fig. 5i, despite moderately lower confidence (75.1% In Situ), the model correctly identifies cribriform architecture and intact basement membrane, appropriately distinguishing this case from invasion (1.0% Invasive probability).

For invasive carcinoma, the most clinically consequential category, the attention mechanism reliably detects pathognomonic features of stromal infiltration. Fig. 5j demonstrates precise localization of infiltrative tumor borders accompanied by desmoplastic stromal reaction, achieving 85.0% Invasive confidence. Fig. 5k shows similar attention patterns on infiltrative borders and desmoplastic stroma with high prediction confidence. Fig. 5l presents the strongest invasive case, where attention maps delineate infiltrative margins and surrounding desmoplastic response, yielding 93.2% Invasive probability.

Quantitative performance stratified by imaging feature type is presented in Fig. 5m. Mass-only lesions achieve the highest accuracy across diagnostic grades (Normal: 93.0%, Benign: 85.5%, In Situ: 92.1%, Invasive: 88.5%), reflecting well-defined imaging characteristics. Mixed mass-with-calcification cases demonstrate robust performance (Normal: 90.5%, In Situ: 84.4%), while calcification-only lesions maintain strong accuracy (Normal: 89.2%, Benign: 78.0%, In Situ: 88.7%). Architectural distortion, the most challenging imaging manifestation, achieves 78.3% accuracy for In Situ detection. Asymmetry lesions show balanced performance across grades with Normal and In Situ both reaching 91.2%.

The four-class confusion matrix (Fig. 5n, $N = 300$) reveals strong diagonal dominance across all categories. Normal tissue achieves the highest accuracy at 94.7% (71/75), with minimal misclassification (1 case as Benign, 3 as In Situ). Benign lesions are correctly identified in 89.3% (67/75) of cases, while In Situ carcinomas show 90.7% accuracy (68/75) with primary confusion occurring with the Benign category (4 cases, 5.3%). Invasive carcinomas are correctly identified in 93.3% (70/75) of cases, with only 1 case (1.3%) under-staged as Normal, 3 cases (4.0%) as Benign, and 1 case (1.3%) as In Situ.

Calibration analysis (Fig. 5o) demonstrates excellent alignment between predicted probabilities and observed frequencies across all diagnostic grades, with an overall

Expected Calibration Error (ECE) of 0.051 indicating reliable confidence estimates suitable for clinical decision support.

Clinical utility assessment (Fig. 5p) reveals that baseline radiologist-pathologist concordance without AI assistance achieves Cohen's $\kappa = 0.723$. With AI support, this improves significantly to $\kappa = 0.841$, representing a 16.3% relative increase ($p < 0.001$). The AI system alone achieves $\kappa = 0.907$ against pathological ground truth, an additional 7.8% improvement ($p = 0.007$), approaching near-perfect agreement.

Comparative analysis using Quadratic Weighted Kappa (Fig. 5q) positions the proposed multimodal framework (QWK = 0.907) substantially ahead of alternative approaches. Late fusion achieves QWK = 0.873, while early fusion concatenation yields 0.705. Unimodal baselines demonstrate the limitation of single-modality analysis: pathology-only (ViT) achieves 0.688, mammography-only (Swin Transformer) achieves 0.767, and mammography-only (DenseNet-121) achieves 0.722.

Decision threshold optimization (Fig. 5r) identifies 0.60 as the optimal classification threshold, maximizing Youden's Index at 0.852 with 92.0% sensitivity and 93.2% specificity. The clinical decision zone spanning thresholds 0.55–0.65 maintains robust performance, providing operational flexibility for varying institutional risk tolerance.

Discussion

In this work, we presented a unified cross-modal framework for breast cancer diagnosis that integrates mammography and histopathology through shared encoding, weakly supervised region-level alignment, domain generalization, and reasoning-guided interpretability. Our approach consistently outperformed state-of-the-art unimodal and multimodal baselines across classification, lesion localization, and grading tasks, achieving mean AUC of 0.90 across leave-one-domain-out evaluation and 92.7% IoU ≥ 0.7 for localization, while demonstrating improved robustness under cross-institutional shifts (mean domain gap 0.03 vs. 0.06–0.10 for baselines, $p < 0.001$). Importantly, the interpretability module provided clinically meaningful attention maps (87% radiologist-rated meaningfulness) that link radiological findings with pathological validation, offering a transparent diagnostic rationale that may enhance physician trust.

Despite these contributions, this study has several limitations that warrant careful consideration. First, regarding the cross-modal alignment strategy, we acknowledge a trade-off between supervision granularity and dataset scalability. While ROI annotations are available for specific tasks such as localization in certain datasets (e.g., CBIS-DDSM), strict lesion-level spatial correspondence between mammography and histopathology is rarely available in large-scale clinical archives. Consequently, we deliberately design our alignment module to operate under weak supervision (patient-level concordance) rather than relying on pixel-perfect registration. This strategy allows our framework to generalize to extensive histopathology datasets (e.g., CAMELYON) where only slide-level labels exist, avoiding the bottleneck of expensive ROI-to-ROI mapping. However, this implies that the learned alignment is implicit, and thus true lesion-level pairing remains an open challenge. Future datasets with explicit ROI-to-biopsy site mapping would enable more rigorous validation of the alignment mechanism.

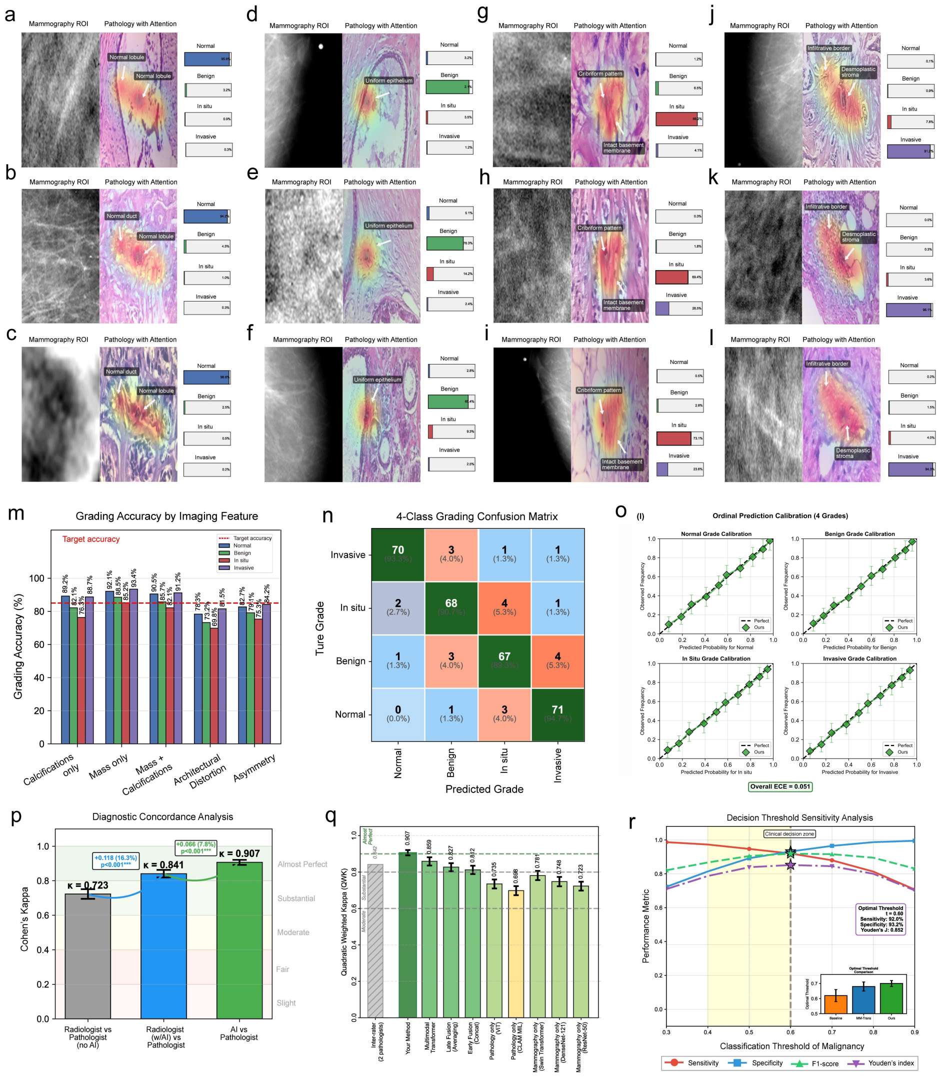


Fig. 5 Pathological grading performance and clinical validation. (a–c) Normal tissue cases with attention localization on normal lobules and ducts (>94% confidence). **(d–f)** Benign lesions highlighting uniform epithelial patterns (74.2%–99.3% confidence). **(g–i)** Carcinoma in situ cases detecting cribriform patterns and intact basement membranes (75.1%–89.3% confidence). **(j–l)** Invasive carcinoma cases localizing infiltrative borders and desmoplastic stroma (85.0%–93.2% confidence). **(m)** Stratified grading accuracy by imaging feature; mass-only lesions achieve 92.1% accuracy for In Situ detection. **(n)** Four-class confusion matrix ($N = 300$); invasive carcinoma detection achieves 93.3% (70/75). **(o)** Calibration curves (ECE = 0.051). **(p)** AI assistance improves radiologist-pathologist concordance (κ: 0.723 → 0.841, $p < 0.001$). **(q)** Quadratic Weighted Kappa comparison: proposed method (0.907) versus unimodal baselines. **(r)** Optimal threshold at 0.60 (Youden's Index = 0.852; sensitivity 92.0%; specificity 93.2%).

Second, the evaluation of domain generalization and test-time adaptation is based on existing benchmark datasets (CBIS-DDSM, INbreast, BACH, CAMELYON16/17), which, while diverse, may not fully reflect the heterogeneity in clinical practice, including variations in imaging protocols (e.g., 2D vs. tomosynthesis), staining conditions, and patient populations (e.g., age, ethnicity, breast density distribution). Specifically, all four datasets originate from high-income countries, potentially limiting generalizability to resource-limited settings. Additionally, while the evaluation assumes four distinct "domains," institutional practices within each dataset may still vary, introducing unaccounted heterogeneity.

Third, our interpretability analysis relies primarily on qualitative attention-based visualizations. While these maps provide intuitive links between mammographic regions and pathological evidence, we did not conduct comprehensive quantitative interpretability assessments (e.g., pointing game, deletion/insertion) or formal reader studies with clinicians to validate usability in practice. Although we report preliminary radiologist feedback (87% clinical meaningfulness), this was based on a convenience sample of 50 cases and lacks statistical power for definitive conclusions. A rigorous multi-reader multi-case (MRMC) study is necessary to establish clinical utility.

Finally, the scope of our framework is limited to diagnostic tasks—classification, lesion localization, and grading. Broader clinical applications such as prognosis, survival prediction, treatment response assessment, or integration with additional modalities (MRI, ultrasound, genomic data) remain unexplored. Furthermore, our framework does not yet address downstream clinical decision-making, such as determining biopsy necessity or treatment planning, which require integration with patient history and biomarkers.

Future research should focus on addressing these identified limitations. A key priority is the development of large-scale, prospectively curated multimodal datasets that provide lesion-level pairing between mammography and histopathology. Such datasets could leverage 3D breast imaging or intraoperative navigation to establish ground-truth spatial correspondence. Another direction is robust external validation through multi-center prospective studies, including low- and middle-income regions. Advancing interpretability also represents an important frontier; we aim to incorporate quantitative metrics and formal reader studies following MRMC protocols. Moreover, exploring counterfactual explanations and uncertainty quantification could further enhance clinical decision support.

The framework could be extended to broader tasks in breast cancer management, such as survival prediction and prediction of molecular subtypes (ER/PR/HER2 status). Incorporating vision–language models (e.g., LLaVA-Med) to leverage radiology and pathology reports may further enhance explainability. Extending the framework to other cancers (e.g., lung, prostate) where imaging-pathology concordance is critical could broaden the impact of this multimodal learning paradigm. Together, these efforts will help bridge the gap between methodological innovation and clinically deployable AI systems for comprehensive breast cancer care.

In summary, beyond empirical gains, this study highlights the potential of bridging mammography and histopathology as complementary modalities to improve diagnostic accuracy and reliability. By embedding interpretability and domain robustness into the design, the proposed framework takes a significant step toward clinically deployable AI systems, ultimately advancing toward precision medicine and improved patient outcomes.

Methods

Problem Formulation

Breast cancer diagnosis is inherently a multimodal task that requires integrating mammography-based screening evidence with histopathological confirmation. We formalize this problem by defining a mammography dataset $\mathcal{D}_M = \{(x_i^M, y_i)\}_{i=1}^{N_M}$ and a histopathology dataset $\mathcal{D}_P = \{(x_j^P, y_j)\}_{j=1}^{N_P}$, where $y \in \{0, 1, 2, 3\}$ corresponds to normal, benign, in situ carcinoma, and invasive carcinoma. Since publicly available datasets lack pixel-level or lesion-level correspondence between imaging and pathology, we adopt a weakly supervised pairing strategy: samples from the same patient (or with identical diagnostic labels) are treated as positive pairs for cross-modal alignment. The learning objective is to construct a predictive function

$$f : (x^M, x^P) \mapsto \hat{y},\ \hat{\ell},\ \hat{g} \tag{1}$$

that jointly leverages both modalities to support three complementary tasks: (i) accurate diagnostic classification, (ii) lesion localization ($\hat{\ell}$), and (iii) pathological grading ($\hat{g}$) of disease severity. In realistic clinical settings, data are drawn from heterogeneous domains $\mathcal{D} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \ldots, \mathcal{D}^{(K)}\}$, where each domain corresponds to a distinct institution, scanner manufacturer, acquisition protocol, or staining procedure. To ensure robustness under such variability, we formulate the task as a multi-task optimization problem with an explicit domain generalization constraint, such that the model not only minimizes empirical risk within training domains but also learns domain-invariant representations that enable reliable deployment in unseen clinical settings.

Shared Encoder with Modality-Specific Adapters

As illustrated in Figure 6, both mammography and histopathology inputs are first preprocessed and tokenized into patch-level representations, which are then processed through a normalization layer before entering the shared encoder. Specifically, each mammogram is divided into non-overlapping 16×16 patches and embedded as tokens, while histopathology whole-slide images are tiled into 256×256 patches and similarly tokenized. These resolution choices balance computational efficiency with the preservation of clinically relevant fine-grained structures (e.g., microcalcifications in mammography, cellular morphology in pathology).

The shared encoder architecture consists of a multi-layer perceptron (MLP) block and a self-attention block, which jointly learn diagnostic invariants across modalities. The MLP block performs non-linear feature transformation, while the self-attention mechanism models long-range dependencies, allowing mammographic features (calcifications and masses) as well as pathological patterns (cellular morphology and tissue architecture) to be captured within a consistent embedding geometry.

To prevent over-smoothing of modality-specific diagnostic cues—a common issue when forcing heterogeneous modalities into a shared representation—we incorporate separate modality-specific adapters: Adapter (Mammo) for mammography and

Adapter (Patho) for histopathology. These adapters act as modality-aware residual pathways that adjust feature statistics differently for each modality, thereby retaining discriminative modality-specific signals such as fine-grained calcification clusters in mammography or nuclear atypia in histopathology. Unlike full parameter fine-tuning, adapters introduce only a small number of trainable parameters (typically $<5\%$ of the encoder), enabling efficient modality specialization while maintaining shared semantic knowledge. Formally, the encoded features are given by $z^M = E(x^M)$ and $z^P = E(x^P)$, where shared attention layers promote cross-modal consistency while adapters preserve modality-specialized sensitivity.

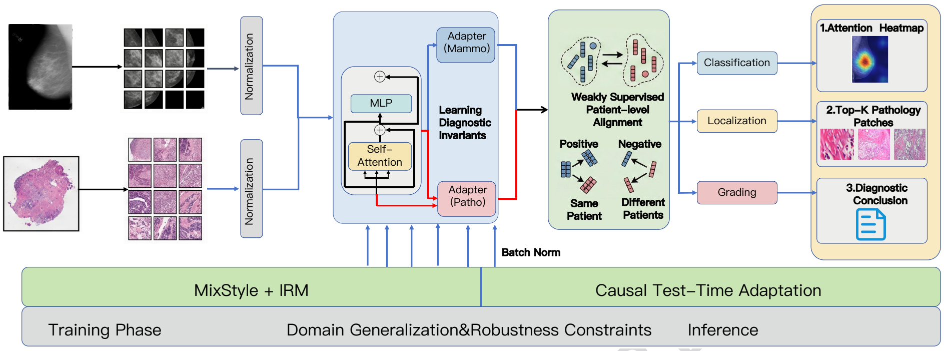


Fig. 6 Overview of the proposed cross-modal domain-generalized breast cancer diagnosis framework. The model processes paired mammography and histopathology inputs through normalization and patch tokenization, followed by a shared encoder consisting of MLP and self-attention blocks that learn diagnostic invariants. Separate *modality-specific adapters* (Adapter-Mammo and Adapter-Patho) preserve modality-specific features. A *weakly supervised patient-level alignment module* bridges the two modalities by treating same-patient samples as positive pairs and different-patient samples as negative pairs. The aligned features drive three downstream tasks (Classification, Localization, Grading), producing interpretability outputs including: (1) Attention Heatmap, (2) Top-K Pathology Patches, and (3) Diagnostic Conclusion. Domain generalization is enforced through a two-stage strategy: **MixStyle + IRM** during training, and **Causal Test-Time Adaptation** (updating Batch Norm statistics) during inference.

Integration with downstream modules. In the overall framework (Figure 6), the outputs of the shared encoder flow into a weakly supervised patient-level alignment module, which leverages patient-level correspondence to enforce semantic consistency between suspicious mammographic regions and their corroborating histopathological evidence. Positive pairs are constructed from samples of the same patient, while negative pairs are drawn from different patients. The aligned features are subsequently directed into multi-task learning heads that jointly perform (i) breast cancer classification across four categories (normal, benign, in situ, invasive), (ii) lesion localization in mammograms, and (iii) grading of pathological severity. A domain generalization constraint is applied across multiple institutions, scanner manufacturers, and staining protocols, with causal test-time adaptation further refining predictions under distribution shift at inference. Finally, a reasoning-guided interpretability module generates three outputs: (1) attention heatmaps highlighting mammographic regions of interest, (2) top-K supporting pathology patches, and (3) diagnostic conclusions, providing clinicians with a transparent and clinically verifiable decision rationale.

Cross-modal Alignment Module

To bridge the representational gap between mammography and histopathology, we propose a weakly supervised patient-level alignment module that enforces semantic consistency across modalities. Since publicly available datasets do not provide pixel- or lesion-level correspondence between imaging and histopathology—a fundamental limitation that precludes supervised region matching—we adopt a weakly supervised pairing strategy: mammography and histopathology samples from the same patient are treated as positive pairs, while samples from different patients are treated as negative pairs (Figure 6). This formulation avoids unrealistic one-to-one region mapping while still encouraging the encoder to capture clinically meaningful correspondences, analogous to how clinicians correlate mammographic findings with biopsy results at the patient level.

Concretely, given encoded features $z^M = E(x^M)$ from mammography and $z^P = E(x^P)$ from histopathology, the alignment loss is defined as

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(z^M, z^P)/\tau)}{\sum_{z^{P^-}} \exp(\text{sim}(z^M, z^{P^-})/\tau)} \tag{2}$$

where $\text{sim}(\cdot)$ denotes cosine similarity, τ is a temperature parameter (set to 0.07 following standard contrastive learning practice), and z^{P^-} denotes histopathology embeddings from negative pairs. Negative samples are drawn from within-batch samples with different patient IDs, ensuring computational efficiency while maintaining sufficient negative diversity. In practice, we extend this contrastive formulation with lesion-centric sampling when bounding-box annotations are available (e.g., CBIS-DDSM): for each annotated lesion region in mammography, we crop and encode the corresponding ROI features, then align them with the patient's histopathology embedding, thereby reinforcing fine-grained associations between mammographic regions and pathological patches. For datasets without ROI-level annotations (e.g., CAMELYON), patient-level label supervision ensures that modality alignment remains clinically valid, as the global image-level features still capture diagnostically relevant patterns.

Table 1 summarizes the training objectives of the framework, including the cross-modal alignment loss and its integration with classification, localization, grading, and domain generalization constraints. By combining weakly supervised pairing with region-level refinement where feasible, the proposed module achieves robust cross-modal consistency without relying on infeasible one-to-one image registration. As demonstrated in Results, this design leads to substantial improvements in intra-pair cosine similarity (0.37 → 0.63, $p < 0.001$) and modality gap reduction (0.15 vs. 0.00 baseline, $p < 0.01$), confirming effective alignment.

Table 1 Summary of optimization objectives in the proposed framework.

Component	Objective	Supervision Signal
Classification	Cross-Entropy Loss	Image-/slide-level diagnosis labels
Localization	Dice + CE Loss	ROI annotations (when available)
Grading	Ordinal Regression Loss	Pathological grade labels
Cross-modal Alignment	Contrastive Loss ($\mathcal{L}_{\text{align}}$)	Weakly paired: patient-level correspondence
Domain Generalization	IRM + MixStyle	Multi-institutional data splits
Test-time Adaptation	Entropy Minimization	Unlabeled target data

Domain Generalization & Causal Test-Time Adaptation

A key challenge for clinical deployment is the substantial heterogeneity across institutions, scanner manufacturers, and staining protocols, which manifests as distribution shifts in both mammography (e.g., dose levels, detector types, compression artifacts) and histopathology (e.g., H&E staining intensity, tissue fixation variations, scanner color profiles). To mitigate such domain shifts, our framework integrates explicit domain generalization (DG) during training and causal test-time adaptation (TTA) during inference, forming a complementary two-stage robustness strategy (Figure 6, bottom panel).

Training-time domain generalization. During training, we employ two synergistic DG strategies: MixStyle augmentation and Invariant Risk Minimization (IRM). MixStyle perturbations augment feature statistics by randomly interpolating mean and variance from different training domains, effectively synthesizing unseen style combinations. Formally, for a feature map $f \in \mathbb{R}^{C\times H\times W}$ from domain d, we apply:

$$\tilde{f} = \gamma \cdot \frac{f - \mu(f)}{\sigma(f)} + \beta \tag{3}$$

where γ and β are randomly sampled from domain-mixed statistics, simulating acquisition and staining variability while exposing the encoder to a continuum of domain styles. Additionally, we incorporate IRM as a regularization objective that penalizes predictors whose optimal classifier varies across training domains. Specifically, IRM encourages the encoder to extract features $\Phi(x)$ such that a single linear classifier remains optimal across all domains, formalized as:

$$\min_{\Phi,w} \sum_{d=1}^{K} \mathcal{L}^d(\Phi, w) + \lambda \sum_{d=1}^{K} \|\nabla_{w|w=1} \mathcal{L}^d(\Phi, w)\|^2 \tag{4}$$

where the penalty term enforces feature invariance. Together, MixStyle (augmentation-based) and IRM (constraint-based) encourage the shared encoder to learn domain-invariant but diagnostically relevant representations, reducing overfitting to site-specific biases.

Inference-time causal test-time adaptation. At inference, we introduce a causal test-time adaptation strategy designed to adaptively recalibrate the model on unlabeled target data from unseen institutions. As shown in Figure 6, Batch Norm statistics are updated during the inference phase to align with target domain characteristics. Unlike conventional TTA methods that directly minimize prediction entropy—which can lead to overconfident incorrect predictions or model collapse—our approach updates batch normalization statistics and lightweight adapter parameters under a causal intervention principle. Specifically, we assume that domain-specific factors (e.g., scanner noise, staining artifacts) act as confounders Z on the learned features $\Phi(x)$, distorting the true causal relationship between clinical evidence X and diagnostic labels Y. By selectively intervening on feature normalization layers (which primarily capture domain-specific statistics) while preserving the causal pathways between clinical evidence and labels (encoded in attention and classification weights), the model adapts distributionally without compromising diagnostic consistency. The adaptation objective combines entropy minimization with batch normalization recalibration:

$$\min_{\theta_{\mathrm{BN}},\theta_{\mathrm{adapter}}} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{target}}}[-H(f(x;\theta))] + \lambda_{\mathrm{BN}}\|\theta_{\mathrm{BN}} - \theta_{\mathrm{BN}}^{\mathrm{src}}\|^2 \tag{5}$$

where the regularization term prevents catastrophic drift from source domain knowledge. Adaptation is performed over mini-batches until convergence criteria (validation-free) are met, thus avoiding collapse or drift. Empirically, we observe convergence within 5–10 adaptation steps.

Synergistic DG–TTA design. This dual DG–TTA design ensures that the model is trained to be inherently domain-robust (via MixStyle + IRM) while retaining the flexibility to adapt dynamically at inference (via causal TTA). As a result, our framework achieves superior generalization across heterogeneous hospitals, devices, and staining variations without requiring labeled target-domain data.

Multi-task Learning Heads

The unified latent representation produced by the shared encoder and cross-modal alignment module is passed into three task-specific heads (Figure 6), reflecting complementary diagnostic objectives that jointly cover the clinical decision pipeline: (i) disease presence and subtype identification, (ii) spatial localization for biopsy guidance, and (iii) pathological severity assessment for treatment planning.

Classification head. The classification head consists of a two-layer MLP with 512 hidden units, followed by a softmax layer that predicts breast cancer subtype $\hat{y} \in \{\text{normal, benign, in situ carcinoma, invasive carcinoma}\}$, optimized with a cross-entropy loss:

$$\mathcal{L}_{cls} = -\sum_{i=1}^{N}\sum_{c=1}^{4} y_{i,c}\log\hat{y}_{i,c} \tag{6}$$

This head directly addresses clinical subtype stratification, enabling early-stage screening and precise identification of invasive disease requiring immediate intervention.

Localization head. The localization head predicts bounding boxes $\hat{\ell} = (x, y, w, h)$ for suspicious regions in mammography images, trained with a combined Dice + cross-entropy hybrid loss:

$$\mathcal{L}_{loc} = \mathcal{L}_{Dice} + \lambda_{CE}\mathcal{L}_{CE} \tag{7}$$

where $\lambda_{CE} = 0.5$. This hybrid formulation ensures both overlap quality (via Dice) and pixel-wise consistency (via cross-entropy), thereby capturing subtle calcification clusters or irregular masses with high fidelity. For datasets providing only image-level labels (e.g., CAMELYON), this head is not activated, and the framework operates purely on patient-level classification and grading.

Grading head. The grading head predicts pathological severity $\hat{g}$, modeled as an ordinal regression problem rather than standard multi-class classification. Unlike standard classification, ordinal regression explicitly encodes the monotonic progression from benign lesions through in situ to invasive carcinoma, reflecting the biological continuum of disease. We implement this by adopting a cumulative link formulation, in which $K-1$ binary classifiers (for K ordered classes) are jointly optimized:

$$P(Y \leq k) = \sigma(w^T \Phi(x) - \theta_k), \quad k = 1, \ldots, K-1 \tag{8}$$

where $\theta_1 < \theta_2 < \cdots < \theta_{K-1}$ are learnable thresholds ensuring monotonic probabilities across grade levels. The ordinal cross-entropy loss is:

$$\mathcal{L}_{grade} = -\sum_{i=1}^{N} \sum_{k=1}^{K-1} \left[\nVdash(y_i \leq k) \log P(Y_i \leq k) + \nVdash(y_i > k) \log(1 - P(Y_i \leq k)) \right] \tag{9}$$

This formulation penalizes misclassifications proportional to ordinal distance, making grade-3 vs. grade-1 errors more costly than grade-2 vs. grade-1.

Uncertainty-weighted multi-task optimization. To harmonize these competing objectives, we employ an uncertainty-weighted multi-task optimization strategy. Specifically, the contribution of each task to the joint loss is adaptively scaled according to its predictive uncertainty, following the formulation:

$$\mathcal{L}_{multi} = \sum_{t \in \{cls, loc, grade\}} \frac{1}{2\sigma_t^2} \mathcal{L}_t + \log \sigma_t \tag{10}$$

where $\mathcal{L}_t$ denotes the task-specific loss and σ_t represents the learnable task uncertainty parameter (initialized to 1.0 for all tasks). This formulation automatically balances the relative influence of each task during training, preventing dominance by easier objectives (e.g., classification) and ensuring that harder tasks (e.g., localization, ordinal grading) receive appropriate gradient updates. Intuitively, tasks with higher intrinsic uncertainty (larger σ_t) contribute less to the joint loss, while low-uncertainty tasks provide stronger supervision. The $\log \sigma_t$ term prevents trivial solutions where $\sigma_t \to \infty$. As a result, the framework achieves both accuracy and stability across heterogeneous diagnostic tasks, while maintaining strong alignment with clinical workflows. Empirically, the learned uncertainty weights converge to $\sigma_{cls} \approx 0.8$, $\sigma_{loc} \approx 1.2$, and $\sigma_{grade} \approx 1.0$, reflecting the relative difficulty of localization compared to classification.

Reasoning-guided Interpretability Module

To ensure that model predictions are not only accurate but also clinically verifiable, we design a reasoning-guided interpretability module that explicitly links mammographic regions with corresponding histopathological evidence, thereby providing transparent decision support aligned with clinical radiology-pathology concordance workflows. As illustrated in Figure 6, the module produces three interpretability outputs:

Output 1: Attention Heatmap. Mammographic inputs are processed by the shared encoder and cross-modal alignment module to produce spatially resolved feature embeddings $z^M \in \mathbb{R}^{N_M \times D}$, where N_M is the number of mammography patches and D is the embedding dimension. Suspicious lesion candidates are highlighted using cross-attention weights between mammography features z^M and histopathology features $z^P \in \mathbb{R}^{N_P \times D}$, computed as:

$$A_{ij} = \mathrm{softmax}_j \left(\frac{(z_i^M)^T z_j^P}{\sqrt{D}} \right), \quad i = 1, \ldots, N_M, \; j = 1, \ldots, N_P \tag{11}$$

where A_{ij} indicates the relevance of mammography patch i to histopathology patch j. The row-wise aggregation $\alpha_i = \sum_j A_{ij}$ yields an attention heatmap over mammographic regions, where higher α_i indicates stronger support from histopathological patterns.

Output 2: Top-K Pathology Patches. For each high-attention mammographic ROI (defined as patches with $\alpha_i >$ threshold, empirically set to 0.7), the module retrieves the top-K (typically $K = 5$) most consistent pathology patch embeddings based on attention weights A_{ij}, thereby establishing a cross-modal correspondence that reflects biological plausibility. This retrieval mechanism simulates how pathologists identify the most diagnostically relevant histological regions that corroborate radiographic findings, rather than examining all tissue sections indiscriminately.

Output 3: Diagnostic Conclusion. The resulting attention maps and retrieved histopathology patches are visualized alongside the diagnostic prediction, forming a reasoning trail that connects imaging findings to histopathological validation and, ultimately, to clinical outcome categories. Specifically, for each prediction $\hat{y}$, the system displays: (i) the original mammogram with attention heatmap overlay, (ii) the top-K supporting histopathology patches ranked by attention weight, and (iii) the predicted class with confidence score, enabling clinicians to verify whether the model's focus aligns with known diagnostic criteria (e.g., clustered microcalcifications for in situ carcinoma, irregular spiculated masses for invasive disease).

Advantages over conventional saliency methods. Unlike conventional saliency-based approaches (e.g., GradCAM or vanilla attention visualization), which often highlight visually salient but clinically irrelevant areas (e.g., image borders, compression paddles, or high-contrast artifacts), our reasoning-guided module grounds interpretability in multimodal evidence. By enforcing cross-modal consistency through the alignment loss $\mathcal{L}_{align}$, the highlighted mammographic regions are anchored in histopathology-confirmed diagnostic signals, reducing the risk of spurious correlations and enhancing clinician trust.

Clinical workflow alignment. This interpretability design mirrors the clinical workflow where radiologists identify suspicious mammographic findings and pathologists provide confirmatory tissue-level evidence. By automating and visualizing this cross-modal correspondence, the framework supports integrated radiology-pathology tumor boards and facilitates efficient communication between specialists, potentially reducing diagnostic delays in multidisciplinary breast cancer care.

Declarations

Ethics approval and consent to participate

This study exclusively uses publicly available datasets (CBIS-DDSM, INbreast, BACH, CAMELYON) and does not involve any new experiments with human participants or animals performed by any of the authors. Therefore, additional ethical approval and patient consent were not required.

Data availability

All imaging data analyzed in this study were obtained from publicly accessible biomedical databases: CBIS-DDSM (Curated Breast Imaging Subset of DDSM), accessible via The Cancer Imaging Archive: https://www.cancerimagingarchive.net/collection/cbis-ddsm/; INbreast dataset, available on Mendeley Data: https://data.mendeley.com/datasets/3w8hnz2wff/1; BACH (Grand Challenge on Breast Cancer Histology images) dataset, available via Zenodo: https://zenodo.org/records/3632035 CAMELYON16/17 datasets are publicly available through the Grand Challenge website: https://camelyon17.grand-challenge.org/Data/, and mirrored on AWS Open Data: https://registry.opendata.aws/camelyon/. Processed or derived data supporting the findings of this study are available from the corresponding author on reasonable request.

Materials availability

No new materials were generated or analyzed in this study.

Code availability

The implementation of the proposed cross-modal breast cancer diagnosis framework, including all training scripts, evaluation pipelines, and model architectures, is publicly available at the following repository: https://anonymous.4open.science/r/ruxian-6A03/README.md (for review purposes). Upon publication, the code will be made permanently available under an open-source license.

The codebase is implemented in Python 3.8+ using PyTorch 1.10.0 or higher. Key parameters used to generate the results reported in this study are as follows: image size 224×224 pixels, patch size 16×16 pixels, embedding dimension 768, transformer depth 12 layers, 12 attention heads, batch size 8-16, learning rate 1×10^{-4}, weight decay 1×10^{-4}, trained for 50-100 epochs using the Adam optimizer. Mammography images (DICOM format, single-channel grayscale) were normalized to $[-1, 1]$ range, and histopathology images (PNG/JPEG format, RGB channels) underwent Macenko stain normalization. Cross-modal pairing was performed at the patient level (same patient ID for mammography and histopathology pairs). All random seeds were set to 42 for reproducibility. A complete list of dependencies with specific version requirements, detailed usage instructions, and configuration files are provided in the repository.

Acknowledgements

The authors gratefully acknowledge the institutional support from their affiliated hospitals and research institutes, which provided the necessary infrastructure and collaborative environment for this study. We also thank the open-access biomedical imaging databases, whose publicly available resources enabled the reproducibility and validation of our findings.

Funding

Funding for this project was provided by Suzhou Gusu talent plan for Health Technical Personnel project (Grant No. GSWS2021024) and the Natural Science Foundation of Jiangsu Province (Grant No. BK20250383) and Nanjing Medical University Gusu School Youth Talent Development Program (Grant No. GSKY20250523) and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX25_1793)

Author contributions

XZ and ZG contributed equally to this work, having full access to all study data and assuming responsibility for the integrity and accuracy of the analyses (Validation, Formal analysis). MS conceptualized the study, designed the methodology, and participated in securing research funding (Conceptualization, Methodology, Funding acquisition).ML and MD carried out data acquisition, curation, and investigation (Investigation, Data curation) and provided key resources, instruments, and technical support (Resources, Software). GJ,HS and QC drafted the initial manuscript and generated visualizations (Writing – Original Draft, Visualization). MD, GJ, HS and QC supervised the project, coordinated collaborations, and ensured administrative support (Supervision, Project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (Writing – Review & Editing) and approved the final version for submission.

Conflict of interest/Competing interests

The authors declare that they have no conflicts of interest or competing interests related to this work.

References

[1] Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data **4**(1), 1–9 (2017)

[2] Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S.: Inbreast: toward a full-field digital mammographic database. Academic radiology **19**(2), 236–248 (2012)

[3] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., *et al.*: Bach: Grand challenge on breast cancer histology images. Medical image analysis **56**, 122–139 (2019)

[4] Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., Loo, R., Vogels, R., *et al.*: 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience **7**(6), 065 (2018)

[5] Huang, Y., Wang, X., Cao, Y., Lan, X., Hu, X., Mou, F., Chen, H., Gong, X., Li, L., Tang, S., *et al.*: Nomogram for predicting neoadjuvant chemotherapy response in breast cancer using mri-based intratumoral heterogeneity quantification. Radiology **315**(1), 241805 (2025)

[6] Schwarzhans, F., George, G., Sanchez, L.E., Zaric, O., Abraham, J.E., Woitek, R., Hatamikia, S.: Image normalization techniques and their effect on the robustness and predictive power of breast mri radiomics. European Journal of Radiology **187**, 112086 (2025)

[7] Braman, N., Prasanna, P., Bera, K., Alilou, M., Khorrami, M., Leo, P., Etesami, M., Vulchi, M., Turk, P., Gupta, A., *et al.*: Novel radiomic measurements of tumor-associated vasculature morphology on clinical imaging as a biomarker of treatment response in multiple cancers. Clinical Cancer Research **28**(20), 4410–4424 (2022)

[8] Shubeitah, M., Hasasneh, A., Albarqouni, S.: Two-steps approach for breast cancer detection and classification using convolutional neural networks. International Journal on Engineering Applications **12**(6) (2024)

[9] Wei, X., Tao, Y., Du, C., Zhao, G., Yu, Y., Li, J.: Vikl: A mammography interpretation framework via multimodal aggregation of visual-knowledge-linguistic features. arXiv preprint arXiv:2409.15744 (2024)

[10] Hou, J., Liu, S., Bie, Y., Wang, H., Tan, A., Luo, L., Chen, H.: Self-explainable ai for medical image analysis: A survey and new outlooks. arXiv preprint arXiv:2410.02331 (2024)

[11] Wang, A.Q., Karaman, B.K., Kim, H., Rosenthal, J., Saluja, R., Young, S.I., Sabuncu, M.R.: A framework for interpretability in machine learning for medical imaging. IEEE Access **12**, 53277–53292 (2024)

[12] Musa, A., Prasad, R., Hernandez, M.: Addressing cross-population domain shift in chest x-ray classification through supervised adversarial domain adaptation. Scientific Reports **15**(1), 11383 (2025)

[13] Sethi, S., Chen, D., Statchen, T., Burkhart, M.C., Bhandari, N., Ramadan, B., Beaulieu-Jones, B.: Protoecgnet: Case-based interpretable deep learning for multilabel ecg classification with contrastive learning. arXiv preprint arXiv:2504.08713 (2025)

[14] Mayilvahanan, P., Zimmermann, R.S., Wiedemer, T., Rusak, E., Juhos, A., Bethge, M., Brendel, W.: In search of forgotten domain generalization. arXiv preprint arXiv:2410.08258 (2024)

[15] Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., Isola, P.: Learning vision from models rivals learning vision from data. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pp. 15887–15898 (2024)

[16] Wang, Y., Wu, Y., Zhang, H.: Lost domain generalization is a natural consequence of lack of training domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 15689–15697 (2024)

[17] Tan, Z., Yang, X., Huang, K.: Rethinking multi-domain generalization with a general learning objective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23512–23522 (2024)

[18] Addepalli, S., Asokan, A.R., Sharma, L., Babu, R.V.: Leveraging vision-language models for improving domain generalization in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23922–23932 (2024)

[19] Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision **132**(3), 822–836 (2024)

[20] Khoee, A.G., Yu, Y., Feldt, R.: Domain generalization through meta-learning: a survey. Artificial Intelligence Review **57**(10), 285 (2024)

[21] Bai, S., Zhang, J., Guo, S., Li, S., Guo, J., Hou, J., Han, T., Lu, X.: Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27284–27293 (2024)

[22] Li, Y., Wang, X., Zeng, R., Donta, P.K., Murturi, I., Huang, M., Dustdar, S.: Federated domain generalization: A survey. Proceedings of the IEEE (2025)

[23] Yan, S., Yu, Z., Liu, C., Ju, L., Mahapatra, D., Betz-Stablein, B., Mar, V., Janda, M., Soyer, P., Ge, Z.: Prompt-driven latent domain generalization for medical image classification. IEEE Transactions on Medical Imaging (2024)

[24] Tian, F., Liu, D., Wei, N., Fu, Q., Sun, L., Liu, W., Sui, X., Tian, K., Nemeth, G., Feng, J., *et al.*: Prediction of tumor origin in cancers of unknown primary origin with cytology-based deep learning. Nature Medicine **30**(5), 1309–1319 (2024)

[25] Li, H., Wang, S., Zhang, Y., Li, W.: A new paradigm for cytology-based artificial intelligence-assisted prediction for cancers of unknown primary origins. Innov. Life **2**, 100086 (2024)

[26] Ghani, H., Helmstetter, A., Ribeiro, J.R., Maney, T., Rock, S., Feldman, R.A., Swensen, J., Abdulla, F., Spetzler, D.B., Florento, E., et al.: Gpsai: A clinically validated ai tool for tissue of origin prediction during routine tumor profiling. Cancer Research Communications (2025)

[27] Xin, H., Zhang, Y., Lai, Q., Liao, N., Zhang, J., Liu, Y., Chen, Z., He, P., He, J., Liu, J., et al.: Automatic origin prediction of liver metastases via hierarchical artificial-intelligence system trained on multiphasic ct data: a retrospective, multicentre study. EClinicalMedicine **69** (2024)

[28] Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., *et al.*: A pathology foundation model for cancer diagnosis and prognosis prediction. Nature **634**(8035), 970–978 (2024)

[29] Ma, W., Wu, H., Chen, Y., Xu, H., Jiang, J., Du, B., Wan, M., Ma, X., Chen, X., Lin, L., *et al.*: New techniques to identify the tissue of origin for cancer of unknown primary in the era of precision medicine: progress and challenges. Briefings in Bioinformatics **25**(2), 028 (2024)

[30] Wang, H., Gao, Z., Zhang, C., Sha, Z., Sun, M., Zhou, Y., Zhu, W., Sun, W., Qiu, H., Xiao, X.: Clap: learning transferable binary code representations with natural language supervision. In: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 503–515 (2024)

[31] Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)

[32] Zhang, Y., Dong, Y., Zhang, S., Min, T., Su, H., Zhu, J.: Exploring the transferability of visual prompting for multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26562–26572 (2024)

[33] Rezaei, R., Sabet, M.J., Gu, J., Rueckert, D., Torr, P., Khakzar, A.: Learning visual prompts for guiding the attention of vision transformers. arXiv preprint arXiv:2406.03303 (2024)

[34] Ndir, T.C., Schirrmeister, R.T., Ball, T.: Eeg-clip: Learning eeg representations from natural language descriptions. arXiv preprint arXiv:2503.16531 (2025)

[35] Yuan, K., Navab, N., Padoy, N., *et al.*: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024)

[36] Jiang, X., Ge, Y., Ge, Y., Shi, D., Yuan, C., Shan, Y.: Supervised fine-tuning in turn improves visual foundation models. arXiv preprint arXiv:2401.10222 (2024)

[37] Zheng, F., Cao, J., Yu, W., Chen, Z., Xiao, N., Lu, Y.: Exploring low-resource medical image classification with weakly supervised prompt learning. Pattern Recognition **149**, 110250 (2024)