GAZEBREAST-FM: AN OMNI BREAST IMAGING FOUNDATION MODEL WITH GAZE-GUIDED DIAGNOSTIC EVIDENCE ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Breast cancer diagnosis relies on complementary information from different breast imaging modalities, yet developing separate models for individual modalities and tasks requires substantial manual annotation and limits representation sharing across imaging domains. Foundation models can reduce this dependence by learning transferable representations through large-scale pre-training, but existing foundation models for breast imaging remain largely modality-specific and are not designed to learn jointly from heterogeneous breast imaging data within a unified framework. GazeBreast-FM is an Omni breast imaging foundation model pre-trained on 302,047 images from 23 datasets. The framework jointly optimizes MAE-style masked reconstruction and CLIP-style image–report alignment, while introducing a spatial attention prior generated from simulated radiologist-like gaze trajectories to allocate sparse diagnostic evidence between masked reconstruction and visible-region semantic learning. A salient-evidence preservation constraint prevents diagnostically relevant information from being completely masked, and diagnostic-evidence coverage is used to dynamically coordinate the two learning objectives. Diagnostic evidence that remains visible after masking is further used for clinical concept learning, while a report-derived Clinical Graph captures continuous clinical relationships across cases and provides structured semantic supervision. Across multiple downstream tasks and breast imaging modalities, GazeBreast-FM achieves state-of-the-art performance, demonstrating transferable representations that support global diagnosis, lesion-level spatial understanding, and fine-grained clinical semantics.https://github.com/aaaaaaa12344/GazeBreastFM

1 INTRODUCTION

Breast cancer is one of the most common malignancies among women worldwide, and breast imaging plays a central role in screening, diagnosis, and treatment assessment. mammogram, ultrasound, and magnetic resonance imaging (MRI) provide distinct yet complementary imaging information. Most existing artificial intelligence systems are developed for a single imaging modality or a specific clinical task, limiting representation reuse across tasks and domains while maintaining a strong dependence on large-scale annotations. Foundation models offer an alternative paradigm by learning transferable representations through large-scale pretraining and supporting adaptation to diverse downstream tasks. Existing breast imaging foundation models, however, remain predominantly modality-specific, with most efforts focusing on mammogram or ultrasound (Ghosh et al., 2024; 2025), leaving unified pretraining across heterogeneous breast imaging modalities relatively underexplored.

Breast imaging encompasses modalities with substantially different imaging characteristics. mammogram primarily depicts breast density, calcifications, and structural abnormalities; ultrasound characterizes lesion morphology, margins, and internal echogenic patterns; and MRI provides complementary information on tissue signal and enhancement characteristics. Despite their distinct visual distributions, these modalities describe partially shared clinical semantics related to lesion morphology, tissue state, and diagnostic attributes. Incorporating heterogeneous breast imaging modalities into a unified pretraining framework can therefore increase both the scale and diversity

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of training data while encouraging the model to learn clinically meaningful representations that are shared across modalities without discarding modality-specific visual characteristics. Recent unified multimodal modeling in breast imaging has also highlighted the value of reducing the fragmentation of models and tasks across individual imaging modalities (Liu et al., 2026b). Following this direction, the present study focuses on mammogram, ultrasound, and MRI and learns transferable breast imaging representations within a unified visual encoding framework.

Masked image modeling and image–text contrastive learning represent two major pretraining paradigms (He et al., 2022; Radford et al., 2021). MAE-style masked reconstruction learns local structure and contextual information by recovering masked image content, whereas CLIP-style image–report learning uses clinical reports as semantic supervision, enabling the visual encoder to capture diagnostically relevant semantics beyond visual appearance alone. Recent masked contrastive approaches in medical imaging have begun to combine these two objectives (Chen et al., 2022; Huang et al., 2024). In breast imaging, diagnostically relevant findings are often spatially sparse. The same diagnostic region can serve as an informative reconstruction target when masked and as direct visual evidence for image–report alignment when kept visible. This study formulates this problem as diagnostic-evidence allocation, in which limited diagnostic information is distributed between masked reconstruction and semantic alignment.

Radiologists exhibit non-uniform visual search and attention patterns during image interpretation (Zhao et al., 2024; Wang et al., 2025). A weak visual-attention prior that approximates radiologist-like reading behavior is therefore introduced to guide the masking and preservation of diagnostically relevant regions. Part of the salient evidence is assigned to masked reconstruction, while the remaining evidence is preserved for image–report alignment. A salient-evidence preservation constraint prevents diagnostically important regions from being completely masked, and diagnostic-evidence coverage is used to adjust the optimization strength of the two pretraining objectives. When sufficient diagnostic information remains visible after masking, the visible diagnostic representation is combined with the global image representation for clinical concept prediction. Case-level image–report alignment introduces clinical semantics into visual representations, but conventional paired-versus-unpaired contrastive learning cannot fully describe partially shared diagnostic information across different cases. Structured concepts and relations extracted from clinical reports are therefore organized into a report-derived Clinical Graph, which models continuous inter-case clinical relationships and provides both shared and modality-specific clinical concept supervision. This structured semantic supervision complements image–report alignment by preserving clinical relationships among cases with similar imaging findings, diagnostic attributes, or pathological states.

The main contributions are summarized as follows:

- An Omni breast imaging foundation model is developed to jointly model mammogram, breast ultrasound, and breast MRI using 302,047 images from 23 datasets.
- A gaze-guided diagnostic-evidence allocation strategy is introduced to distribute sparse diagnostic evidence between MAE-style masked reconstruction and CLIP-style image–report alignment, together with salient-evidence preservation and coverage-aware dynamic weighting to coordinate the two pretraining objectives.
- Preserved visible diagnostic evidence is incorporated into clinical concept learning, while report-derived structured clinical semantics provide inter-case relationships and clinical concept supervision.
- Comprehensive evaluation across three breast imaging modalities, seven downstream tasks, and five evaluation datasets demonstrates state-of-the-art performance on multiple benchmarks.

2 Related Work

Masked modeling and medical vision–language pretraining. Masked image modeling learns transferable visual representations by reconstructing missing content, with MAE demonstrating the effectiveness of high-ratio random masking at scale (He et al., 2022). Medical vision–language methods exploit paired images and clinical reports to learn clinically meaningful representations, progressing from global image–text alignment toward local and multi-granularity semantic correspondence in ConVIRT, GLoRIA, BioViL, MGCA, and AFLoc (Zhang et al., 2022; Huang et al.,

2021; Boecking et al., 2022; Wang et al., 2022a; Yang et al., 2026). Recent approaches increasingly combine visual reconstruction with semantic supervision. M3AE jointly models masked medical images and text, Masked Record and Relation Modeling study learning from incomplete medical information, and MaCo combines masked reconstruction with image–report contrastive learning using correlation-aware weighting (Chen et al., 2022; Zhou et al., 2023; Yang et al., 2023; Huang et al., 2024). These developments motivate a more explicit treatment of how spatially sparse diagnostic evidence is allocated between reconstruction and masked-view semantic learning.

Gaze-informed medical representation learning. Radiologist gaze provides spatial cues about diagnostically relevant image regions. Existing studies use gaze information to improve contrastive-view construction, diagnostic representation learning, weakly supervised localization, segmentation, and vision–language pretraining (Lanfredi et al., 2022; Zhao et al., 2024; Wang et al., 2025; Chen et al., 2026; Liu et al., 2026a). DiffEye models continuous eye movements as an image-conditioned diffusion process and enables scalable generation of diverse visual-search trajectories (Kara et al., 2025). This spatial information is used to coordinate the allocation of diagnostic evidence between masked reconstruction and visible semantic learning, directly supporting gaze-guided diagnostic-evidence allocation.

Structured clinical semantic learning. Clinical cases often share findings and diagnostic attributes that are poorly represented by binary paired-versus-unpaired supervision. MedCLIP introduces knowledge-based semantic matching to model graded clinical similarity, while MedKLIP incorporates structured medical entities and relations into vision–language learning (Wang et al., 2022b; Wu et al., 2023). Image–Graph Contrastive Pretraining further exploits report-derived graph representations, and Ultrasound-CLIP combines structured diagnostic attributes, semantic soft labels, and graph relations to model continuous clinical similarity (Khanna et al., 2023; Jin et al., 2026). Building on this direction, the report-derived Clinical Graph provides continuous inter-case semantics, cross-modality clinical concept supervision, and structured consistency constraints.

3 METHOD

3.1 FRAMEWORK OVERVIEW

Figure 1 summarizes GazeBreast-FM. mammogram, breast ultrasound, and breast MRI share a single visual encoder, through which each image is processed as a clean view and a gaze-guided masked view. The clean view provides a case-level global representation for image–report alignment, whereas the masked view allocates diagnostically salient evidence between masked reconstruction and visible semantic learning. The framework jointly optimizes global image–report alignment, gaze-guided reconstruction, diagnostic semantic alignment, and Clinical Graph consistency. Diagnostic coverage further coordinates the relative strength of reconstruction and semantic alignment during pre-training.

3.2 GAZE-GUIDED DIAGNOSTIC EVIDENCE ALLOCATION

Visual-search prior construction. The open-source DiffEye method (Kara et al., 2025) is adapted separately to mammogram, breast ultrasound, and breast MRI using expert eye-tracking data collected in our laboratory, with 150 images used for adaptation and 15 held-out images for validation in each modality. The modality-adapted models are used to approximate expert visual-search behavior during breast-image interpretation and to generate continuous visual-search trajectories at scale. For image x_i^m of modality $m \in \{\mathrm{MG}, \mathrm{US}, \mathrm{MRI}\}$, where i indexes images, K generated trajectories form $\mathcal{T}_i^m = \{\tau_{ik}^m\}_{k=1}^K$, with k indexing trajectories. After modality-specific filtering, the retained trajectories are aggregated into an image-level attention map A_i^m. Its normalized attention projected onto patch n is denoted by a_{in}^m, and high-response patches form the salient evidence set $\mathcal{S}_i$. The resulting signal is used as a DiffEye-generated weak spatial attention prior rather than as direct expert-gaze annotation for the pre-training corpus.

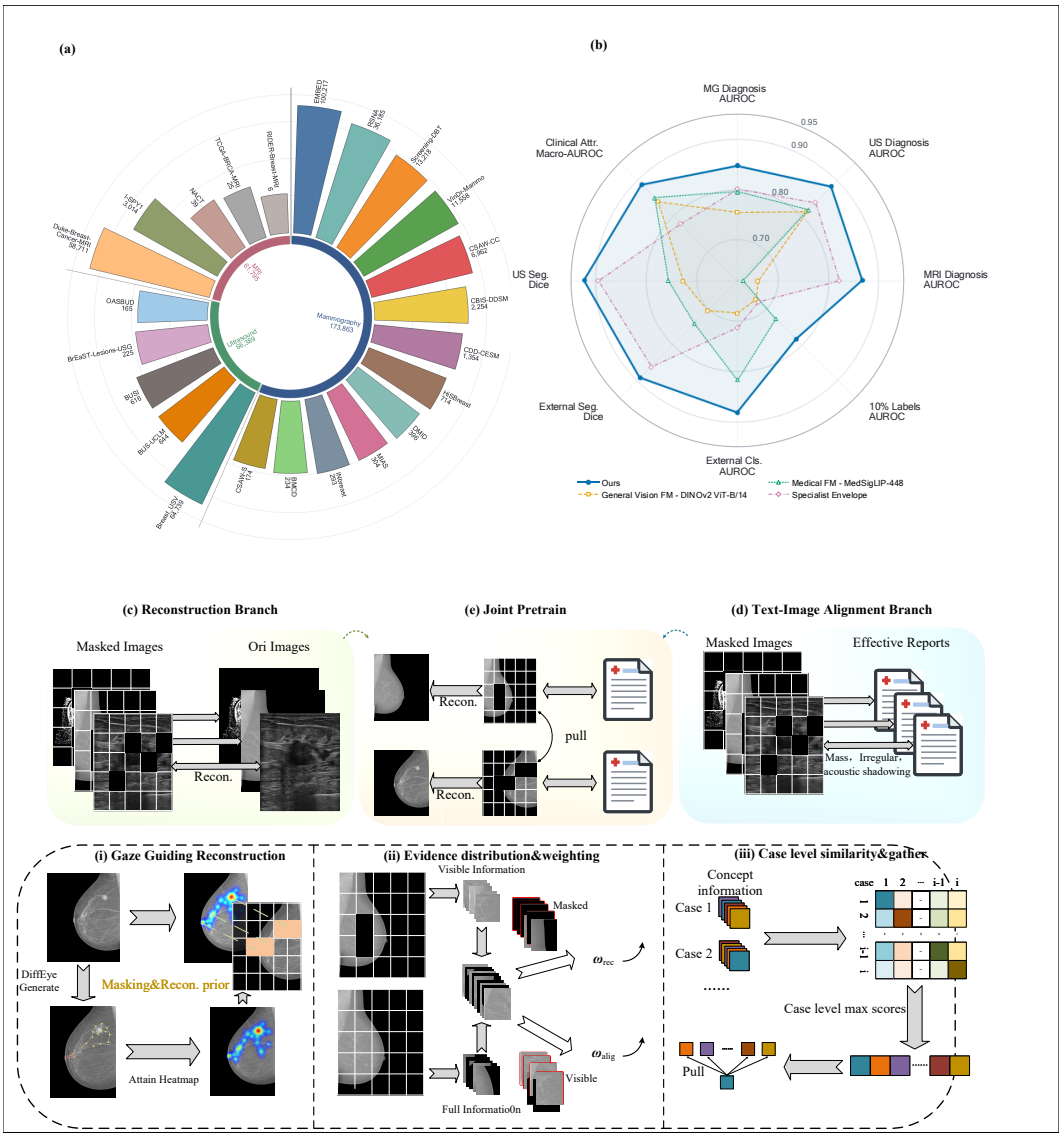


Figure 1: Overview of GazeBreast-FM. (a) Distribution of the breast imaging datasets used for pre-training. (b) Downstream performance across diagnostic, label-efficient, external generalization, spatial, and clinical semantic tasks. The lower panel illustrates the proposed GazeBreast-FM framework.

Attention-guided masking. Let $h_{in} \in \{0, 1\}$ indicate whether patch n belongs to $\mathcal{S}_i$, and let p_{in}^u denote the uniform masking probability. The final masking probability is

$$p_{in}^{\text{mask}} = (1-\gamma)p_{in}^u + \gamma \frac{\left(a_{in}^m + \lambda_h h_{in}\right)^{1/\tau_m}}{\sum_j \left(a_{ij}^m + \lambda_h h_{ij}\right)^{1/\tau_m}}. \tag{1}$$

Here, j indexes patches, λ_h controls the contribution of salient patches, τ_m is the modality-specific sampling temperature, and $\gamma \in [0, 1]$ controls the mixture between attention-guided and uniform sampling.

Diagnostic evidence allocation. Let $M_{in} \in \{0,1\}$ denote the mask state of patch n, where $M_{in} = 1$ indicates masking. For a non-empty salient set, the visible salient-evidence fraction is

$$q_i^{\text{vis}} = \frac{\sum_n \mathbf{1}[n \in \mathcal{S}_i](1 - M_{in})}{|\mathcal{S}_i|} \geq \rho_{\text{vis}}. \tag{2}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function, $|\mathcal{S}_i|$ is the number of salient patches, q_i^{vis} is their visible fraction, and $\rho_{\text{vis}} > 0$ is the minimum visible-salient fraction. The constraint is jointly enforced with the overall masking ratio.

Reconstruction and visible diagnostic evidence. Let $V_{in} \in \{0,1\}$ denote the valid-content indicator, $\hat{u}_{in}^m$ the predicted masked token, $u_{in}^{c,m}$ its clean-view target, and D the token dimension. The operators $\text{LN}(\cdot)$ and $\text{sg}(\cdot)$ denote layer normalization and stop-gradient, respectively. The diagnostic-weighted reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{\sum_{i,n} M_{in} V_{in} (1 + \alpha_r a_{in}^m + \beta_r h_{in}) \frac{\left\| \text{LN}(\hat{u}_{in}^m) - \text{LN}(\text{sg}(u_{in}^{c,m})) \right\|_2^2}{D}}{\sum_{i,n} M_{in} V_{in} (1 + \alpha_r a_{in}^m + \beta_r h_{in}) + \epsilon}. \tag{3}$$

Here, α_r and β_r control attention- and saliency-based reconstruction weighting, respectively, and $\epsilon > 0$ is a numerical stabilizer. Visible valid tokens are pooled using diagnostic attention and salient-patch weighting to obtain the visible diagnostic representation z_i^v of image i.

3.3 Image–Report Alignment

The clean image view produces the global representation z_i^g, while the masked view produces the visible diagnostic representation z_i^v. The corresponding Effective Report is encoded as t_i. Their projected and L_2-normalized representations are denoted by $\tilde{z}_i^g$, $\tilde{z}_i^v$, and $\tilde{t}_i$.

For a batch of B cases, let $r \in \{g, v\}$ select the global or visible visual representation, and let τ denote the contrastive temperature. The symmetric image–report alignment objective is

$$\mathcal{L}_{\text{IR}}^r = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp\left((\tilde{z}_i^r)^\top \tilde{t}_i / \tau\right)}{\sum_j \exp\left((\tilde{z}_i^r)^\top \tilde{t}_j / \tau\right)} + \log \frac{\exp\left(\tilde{t}_i^\top \tilde{z}_i^r / \tau\right)}{\sum_j \exp\left(\tilde{t}_i^\top \tilde{z}_j^r / \tau\right)} \right]. \tag{4}$$

Here, i and j index cases within the batch. The $r = g$ instance defines the global alignment loss $\mathcal{L}_{\text{global}}$, whereas $r = v$ defines the visible alignment loss $\mathcal{L}_{\text{visible}}$.

The normalized global and visible representations are equally fused into the case-level semantic feature f_i, which is used for semantic-relation, clinical-concept, and concept-consistency learning. After report acceptance, each Effective Report is converted into a case-level Clinical Graph that represents structured clinical relationships across cases and provides supervision for semantic soft targets, clinical concept learning, and structural consistency. Graph construction, provenance, and quality control are detailed in Appendix C. Let Q_{ij} denote the sparse semantic target weight between cases i and j, constructed from Effective Reports and report-derived Clinical Graph relations, and let $P_{ij}^{I \to T}$ and $P_{ij}^{T \to I}$ denote the corresponding image-to-text and text-to-image softmax similarity probabilities.

$$\mathcal{L}_{\text{soft}} = -\frac{1}{2B} \sum_{i,j} \left[Q_{ij} \log P_{ij}^{I \to T} + Q_{ji} \log P_{ij}^{T \to I} \right]. \tag{5}$$

The same case-level feature supports shared and modality-specific clinical concept prediction. Let $\mathcal{H}$ denote the set of concept heads with valid supervision, h index a concept head, a_h denote its static weight, and ℓ_h its prediction loss. The clinical concept objective is

$$\mathcal{L}_{\text{concept}} = \frac{\sum_{h \in \mathcal{H}} a_h \ell_h}{\sum_{h \in \mathcal{H}} a_h + \epsilon}. \tag{6}$$

Here, ℓ_h is binary cross-entropy for binary or multilabel concepts and categorical cross-entropy for multiclass concepts. Only concept targets valid for the current imaging modality participate in the loss.

Clinical Graph supervision further provides inter-case and prototype-level concept consistency. Let Ω denote the set of valid case pairs, Γ the set of valid sample–concept–prototype assignments, A_{ij}^{cc} the concept-derived agreement between cases i and j, and p_{hv} the prototype corresponding to value v of concept h. With $\cos(\cdot,\cdot)$ denoting cosine similarity, the concept-consistency loss is

$$\mathcal{L}_{\mathrm{cc}} = \mathrm{mean}_{\mathrm{valid}} \left\{ \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \left[\frac{1+\cos(f_i, f_j)}{2} - A_{ij}^{\mathrm{cc}} \right]^2, \; \frac{1}{|\Gamma|} \sum_{(i,h,v)\in\Gamma} [1 - \cos(f_i, p_{hv})] \right\}. \quad (7)$$

Here, $\mathrm{mean}_{\mathrm{valid}}$ averages the consistency terms that are valid in the current batch.

Valid Clinical Graph relations additionally define the structural consistency loss

$$\mathcal{L}_{\mathrm{graph}} = \frac{\sum_{e\in\mathcal{R}} \rho_e \phi_e}{\sum_{e\in\mathcal{R}} \rho_e + \epsilon}. \quad (8)$$

Here, $\mathcal{R}$ is the set of valid graph relations, e indexes a relation, ρ_e is its semantic weight, and ϕ_e is its relation-specific penalty. Conflict relations use the product of source and target probabilities, implication or risk-order relations use the positive part of their probability difference, and shared-concept relations use the squared probability difference.

3.4 Coverage-Aware Joint Optimization

Let c_m and c_v denote the batch-level proportions of diagnostic attention assigned to masked and visible valid patches, respectively. Given their target coverages κ_m and κ_v, let $b \in \{\mathrm{rec}, \mathrm{alig}\}$ select the reconstruction or text-alignment branch. The corresponding dynamic weight is

$$\omega_b = \begin{cases} \mathrm{clip}\left(\dfrac{c_v}{\kappa_v}, \omega_{\min}, \omega_{\max}\right), & b = \mathrm{alig}, \\ \mathrm{clip}\left(\dfrac{c_m}{\kappa_m}, \omega_{\min}, \omega_{\max}\right), & b = \mathrm{rec},\ c_m > \kappa_m, \\ 1, & b = \mathrm{rec},\ c_m \le \kappa_m. \end{cases} \quad (9)$$

Here, $\omega_{\min}$ and $\omega_{\max}$ are the lower and upper bounds of the dynamic weights, and $\mathrm{clip}(\cdot)$ restricts its input to this interval. The cases $b = \mathrm{rec}$ and $b = \mathrm{alig}$ define ω_{rec} and ω_{alig}, respectively.

The visible alignment, semantic soft-relation, clinical concept, and concept consistency objectives constitute the semantic learning components of the text-alignment branch and are jointly represented by

$$\mathcal{L}_{\mathrm{alig}} = \lambda_{\mathrm{visible}}\mathcal{L}_{\mathrm{visible}} + \lambda_{\mathrm{soft}}\mathcal{L}_{\mathrm{soft}} + \lambda_{\mathrm{concept}}\mathcal{L}_{\mathrm{concept}} + \lambda_{\mathrm{cc}}\mathcal{L}_{\mathrm{cc}}. \quad (10)$$

Here, $\lambda_{\mathrm{visible}}$, λ_{soft}, $\lambda_{\mathrm{concept}}$, and λ_{cc} are the corresponding static loss weights.

The final joint optimization objective is

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{global}}\mathcal{L}_{\mathrm{global}} + \omega_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \omega_{\mathrm{alig}}\mathcal{L}_{\mathrm{alig}} + \lambda_{\mathrm{graph}}\mathcal{L}_{\mathrm{graph}}. \quad (11)$$

Here, $\lambda_{\mathrm{global}}$ and λ_{graph} are the static weights of the global image–report alignment and Clinical Graph structural consistency objectives, respectively.

4 Experiments

4.1 Experimental Setup

GazeBreast-FM initializes its visual and text encoders from Mammo-FM and BiomedCLIP, respectively, and is jointly pre-trained on 302,047 breast images from 23 datasets; dataset composition and provenance are summarized in Appendix A. Downstream benchmarks are separated from the pre-training data and use patient-level splits. Classification uses linear classifiers on frozen image representations, image–text retrieval is evaluated zero-shot in the frozen joint embedding space, and spatial and clinical-attribute tasks use task-specific prediction heads. All trainable downstream experiments use fixed seeds 11, 12, and 13, with checkpoint and threshold selection based only on validation data; the test set is used only for final evaluation. Results are reported in the order of mammogram, ultrasound, and MRI.

Table 1: Breast disease classification across mammogram, ultrasound, and breast MRI. Results are reported as mean ± population standard deviation over three downstream seeds.

Modality	Dataset	Model	AUROC	AUPRC	F1
Mammogram	OMAMA	DINOv2 ViT-B/14	0.7555 ± 0.0016	0.1792 ± 0.0029	0.2329 ± 0.0011
		MedSigLIP-448	0.7955 ± 0.0007	0.2245 ± 0.0019	0.2712 ± 0.0050
		Mammo-FM	0.8009 ± 0.0010	0.2114 ± 0.0005	0.2834 ± 0.0020
		Ours	$\mathbf{0.8467 \pm 0.0028}$	$\mathbf{0.2738 \pm 0.0049}$	$\mathbf{0.3371 \pm 0.0063}$
Ultrasound	BUSI-WHU	DINOv2 ViT-B/14	0.8134 ± 0.0053	0.7220 ± 0.0077	0.6647 ± 0.0090
		MedSigLIP-448	0.8178 ± 0.0011	0.7343 ± 0.0046	0.6780 ± 0.0020
		OpenUS	0.8380 ± 0.0047	0.7747 ± 0.0048	0.6462 ± 0.0107
		Ours	$\mathbf{0.8826 \pm 0.0037}$	$\mathbf{0.8269 \pm 0.0045}$	$\mathbf{0.7358 \pm 0.0072}$
MRI	fastMRI Breast	DINOv2 ViT-B/14	0.6605 ± 0.0048	0.4756 ± 0.0071	0.5344 ± 0.0064
		MedSigLIP-448	0.6311 ± 0.0066	0.3952 ± 0.0046	0.4338 ± 0.0424
		MeisenMeister	0.8210 ± 0.0007	0.6156 ± 0.0011	0.6486 ± 0.0182
		Ours	$\mathbf{0.8674 \pm 0.0056}$	$\mathbf{0.6812 \pm 0.0079}$	$\mathbf{0.7095 \pm 0.0098}$

Table 2: Lesion segmentation and localization across mammogram, ultrasound, and breast MRI. Results are reported as mean ± population standard deviation over three downstream seeds.

Modality	Dataset	Model	Dice ↑	mAP@0.5 ↑
mammogram	BCDR	DINOv2 ViT-B/14	0.5989 ± 0.0033	0.5223 ± 0.0068
		MedSigLIP-448	0.6514 ± 0.0040	0.6745 ± 0.0075
		Mammo-CLIP	0.7216 ± 0.0052	0.7498 ± 0.0059
		GazeBreast-FM	$\mathbf{0.7649 \pm 0.0047}$	$\mathbf{0.7876 \pm 0.0054}$
Ultrasound	BUSI-WHU	DINOv2 ViT-B/14	0.7273 ± 0.0034	0.6392 ± 0.0068
		MedSigLIP-448	0.7567 ± 0.0008	0.7265 ± 0.0093
		OpenUS	0.5895 ± 0.0078	0.5332 ± 0.0171
		GazeBreast-FM	$\mathbf{0.8328 \pm 0.0029}$	$\mathbf{0.8564 \pm 0.0047}$
MRI	AMBL	DINOv2 ViT-B/14	0.5891 ± 0.0029	0.6104 ± 0.0027
		MedSigLIP-448	0.6372 ± 0.0022	0.6145 ± 0.0009
		MAMA-MIA nnU-Net	0.6968 ± 0.0089	0.7337 ± 0.0048
		GazeBreast-FM	$\mathbf{0.7424 \pm 0.0065}$	$\mathbf{0.7926 \pm 0.0042}$

4.2 Breast Disease Classification

Case-level diagnostic transfer is evaluated on OMAMA, BUSI-WHU, and fastMRI Breast for mammogram, breast ultrasound, and MRI, respectively. The three benchmarks assess whether a shared pretrained representation can support diagnostic discrimination across distinct breast imaging modalities.

GazeBreast-FM consistently improves classification across the three modalities, reaching AUROCs of 0.8467, 0.8826, and 0.8674 on OMAMA, BUSI-WHU, and fastMRI Breast, respectively, while also obtaining the highest AUPRC and F1 in each setting. The gains are consistent across general and modality-specific comparators, indicating that the benefit is not confined to a single imaging domain. This trend is also reflected in the feature-space visualization in Fig. 2: the silhouette score increases from -0.023 to 0.137 for mammogram, from 0.025 to 0.486 for ultrasound, and from -0.002 to 0.382 for MRI. Ultrasound and MRI show particularly clear benign–malignant separation, while mammogram exhibits improved class organization despite remaining overlap.

4.3 Lesion Segmentation and Localization

Lesion segmentation and localization evaluate whether the learned representations preserve clinically meaningful spatial information across breast imaging modalities. We conduct evaluation on mammogram, breast ultrasound, and breast MRI, reporting Dice for pixel-level lesion segmenta-

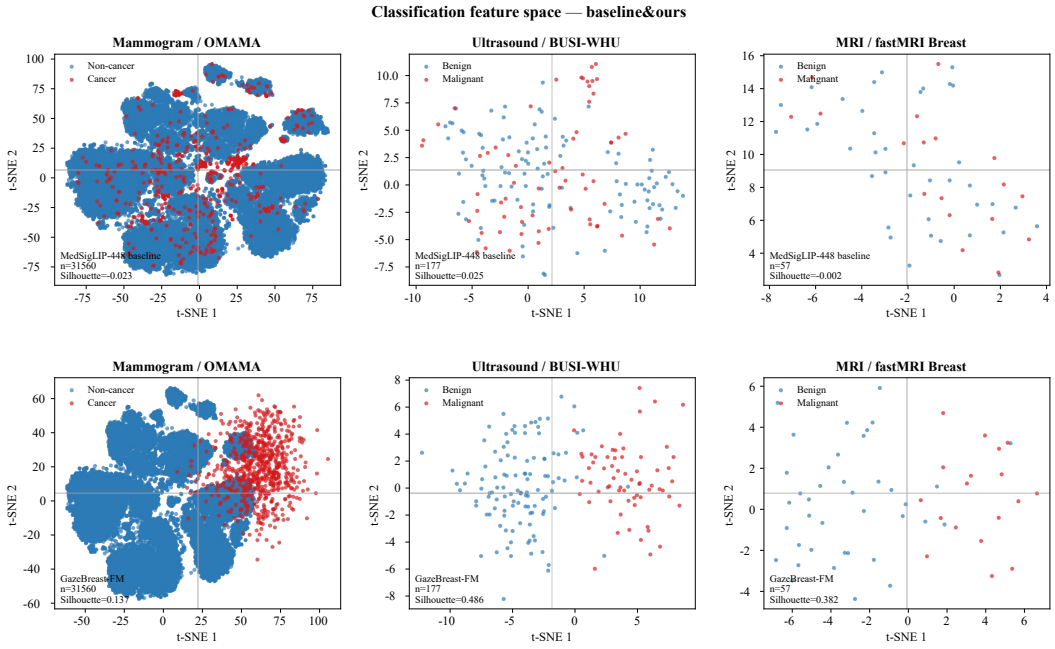


Figure 2: t-SNE visualization of classification representations from MedSigLIP-448 (top) and GazeBreast-FM (bottom) on OMAMA, BUSI-WHU, and fastMRI Breast.

tion together with mAP@0.5 for lesion localization. These complementary metrics assess both the delineation of lesion regions and the identification of their spatial extent.

GazeBreast-FM achieves the strongest spatial performance across mammogram, ultrasound, and MRI, with consistent gains in Dice and mAP@0.5. The agreement between segmentation and localization metrics indicates that the learned representation supports both lesion-region delineation and lesion-level localization across distinct imaging characteristics. This trend is also reflected in Fig. 3: compared with the strong spatial baselines, GazeBreast-FM produces more concentrated lesion-related responses and less diffuse activation in non-lesion regions across all three modalities. Mammographic responses remain focused within complex breast tissue, ultrasound responses more completely cover lesion regions with clearer boundaries, and MRI responses form more continuous activation over the lesion extent. Together, the quantitative and qualitative results indicate robust lesion-selective spatial transfer across the three breast imaging modalities.

4.4 Ablation Studies

We conduct controlled ablations to examine the contributions of the two principal pre-training objectives, gaze guidance, diagnostic evidence allocation, and coverage-aware dynamic weighting.

Table 3: Ablation of the principal components of GazeBreast-FM.

Variant	MG AUROC	MRI AUROC	US Dice	Attribute Macro-AUROC
Full	$\mathbf{0.8467 \pm 0.0028}$	$\mathbf{0.8674 \pm 0.0056}$	$\mathbf{0.8328 \pm 0.0029}$	$\mathbf{0.8876 \pm 0.0016}$
Reconstruction only	0.8142 ± 0.0047	0.8326 ± 0.0068	0.8097 ± 0.0045	0.8438 ± 0.0039
Alignment only	0.8315 ± 0.0039	0.8512 ± 0.0062	0.7894 ± 0.0057	0.8721 ± 0.0028
w/o Gaze Guidance	0.8238 ± 0.0042	0.8427 ± 0.0065	0.7986 ± 0.0051	0.8604 ± 0.0031
w/o Evidence Allocation	0.8356 ± 0.0035	0.8558 ± 0.0060	0.8149 ± 0.0042	0.8752 ± 0.0025
w/o Dynamic Weighting	0.8415 ± 0.0031	0.8619 ± 0.0058	0.8256 ± 0.0036	0.8820 ± 0.0020

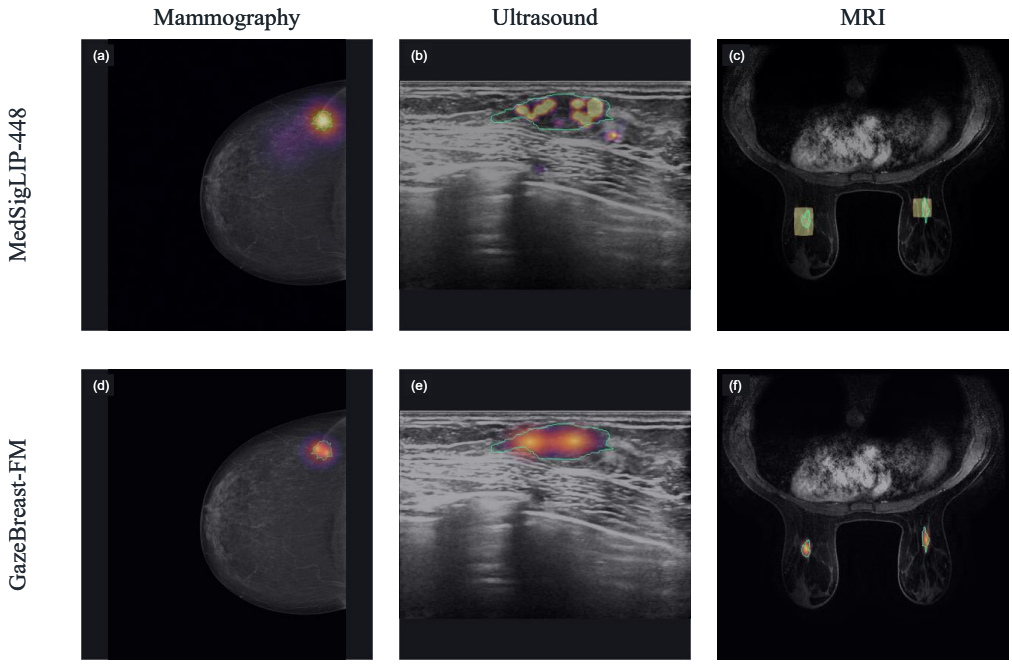


Figure 3: Spatial-response comparison between MedSigLIP-448 and GazeBreast-FM across mammogram, breast ultrasound, and breast MRI. Higher responses indicate regions receiving stronger lesion-related representation activation.

The full model achieves the strongest performance across all evaluation metrics, while both single-objective variants show clear degradation, indicating complementary contributions from reconstruction and alignment-side supervision. Reconstruction-only training retains comparatively stronger spatial transfer but loses more diagnostic and clinical-semantic information, whereas the alignment-only variant preserves stronger semantic performance while showing a larger reduction on the spatial task. Removing gaze guidance produces the largest or near-largest overall degradation, demonstrating the importance of the weak spatial-attention prior to the full learning framework. Performance further decreases when diagnostic evidence allocation is removed, showing that actively coordinating which informative regions are masked and which remain visible contributes beyond gaze-derived weighting alone. Finally, disabling coverage-aware dynamic weighting yields a smaller but consistent decline across tasks, indicating that adaptive coordination between reconstruction and alignment provides additional benefit once diagnostic evidence allocation is established. Overall, the ablation pattern supports the complementary roles of the two learning objectives and the progressive contribution of gaze guidance, evidence allocation, and coverage-aware coordination.

5 Conclusion

GazeBreast-FM is an Omni breast imaging foundation model that uses gaze-guided diagnostic evidence allocation to coordinate masked reconstruction and image–report alignment. Its representations transfer across mammogram, ultrasound, and MRI in diagnostic and lesion-level downstream tasks. A single shared representation may incompletely capture modality-specific characteristics, while region-level correspondence between visual evidence and clinical information remains limited.

AI USE STATEMENT

Generative AI tools and AI-assisted coding systems were used during research development and manuscript preparation to support literature search and summarization, discussion and review of methodological and experimental designs, code implementation and revision, program debugging, analysis and interpretation of experimental results, and manuscript organization, drafting, translation, and language editing. All AI-assisted research designs, implementation plans, code, analyses, and manuscript content were reviewed by the authors and verified against source data, experimental outputs, and reproducible program results. In addition, as part of the data-processing pipeline, a generative medical vision–language model was used to produce case-level reports for cases lacking trustworthy official clinical reports, conditioned on structured clinical facts and the corresponding medical images. Generated reports were subjected to factual-consistency checks and quality-control criteria, and only reports satisfying the predefined acceptance criteria were used for subsequent pre-training. For datasets with trustworthy official reports, generated reports were not used as replacements. Generative AI was not used to generate or modify the experimental performance results reported in this paper. The authors take full responsibility for the final research methodology, implementation, experimental results, analyses, and manuscript content.

ETHICS STATEMENT

This study uses existing breast medical imaging datasets for model pre-training and downstream evaluation and involves no new participant recruitment, clinical intervention, or prospective data collection. Each dataset is handled according to the access conditions and permitted uses specified by its original provider, with data provenance and train–test separation checked before inclusion in the experimental pipeline. Patient-level partitioning is used to prevent data from the same patient from appearing across training, validation, and test sets. The resulting models and experimental findings are intended for research purposes and should not be interpreted as a clinically deployed diagnostic system.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the principal methods, training settings, data splits, and evaluation metrics. Trainable downstream experiments use fixed seeds 11, 12, and 13 and are reported as mean $\pm$ population standard deviation. Patient-level partitioning is used to prevent data leakage. Publicly distributable training configurations, split protocols, and reproducibility materials will accompany the final release.

REFERENCES

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *Computer Vision – ECCV 2022*, pp. 1–21, 2022. doi: 10.1007/978-3-031-20059-5_1.

Jingkun Chen, Haoran Duan, Xiao Zhang, Boyan Gao, Vicente Grau, and Jungong Han. From gaze to insight: Bridging human visual attention and vision language model explanation for weakly-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 45(3):1051–1061, 2026. doi: 10.1109/TMI.2025.3616598.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 679–689, 2022. doi: 10.1007/978-3-031-16443-9_65.

Shantanu Ghosh, Clare B. Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 632–642, 2024. doi: 10.1007/978-3-031-72390-2_59.

Shantanu Ghosh, Vedant Parthesh Joshi, Rayan Syed, Aya Kassem, Abhishek Varshney, Payel Basak, Weicheng Dai, Judy Wawira Gichoya, Hari M. Trivedi, Imon Banerjee, Shyam Visweswaran, Clare B. Poynton, and Kayhan Batmanghelich. Mammo-FM: Breast-specific foundational model for integrated mammographic diagnosis, prognosis, and reporting. *arXiv preprint arXiv:2512.00198*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Weijian Huang, Cheng Li, Hong-Yu Zhou, Hao Yang, Jiarun Liu, Yong Liang, Hairong Zheng, Shaoting Zhang, and Shanshan Wang. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications*, 15:7620, 2024. doi: 10.1038/s41467-024-51749-0.

Jiayun Jin, Haolong Chai, Xueying Huang, Xiaoqing Guo, Zengwei Zheng, Zhan Zhou, Junmei Wang, Xinyu Wang, Jie Liu, and Binbin Zhou. Ultrasound-CLIP: Semantic-aware contrastive pretraining for ultrasound image-text understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6962–6971, 2026.

Ozgur Kara, Harris Nisar, and James M. Rehg. DiffEye: Diffusion-based continuous eye-tracking data generation conditioned on natural images. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Sameer Khanna, Daniel Michael, Marinka Zitnik, and Pranav Rajpurkar. Learning generalized medical image representations through image-graph contrastive pretraining. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 232–243, 2023.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific Data*, 9:350, 2022. doi: 10.1038/s41597-022-01441-z.

Kang Liu, Zhuoqi Ma, Siyu Liang, Yunan Li, Xiyue Gao, Chao Liang, Kun Xie, and Qiguang Miao. Seeing like radiologists: Context- and gaze-guided vision-language pretraining for chest x-rays. *arXiv preprint arXiv:2603.26049*, 2026a.

Yang Liu, Jiajin Zhang, Danyang Tu, Yaojun Hu, Jiao Qu, Jiuyu Zhang, Yu Shi, Wei Fang, Shi Gu, Ling Zhang, and Yingda Xia. BreastGPT: A multimodal large language model for the full spectrum of breast cancer clinical routine. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33536–33549, 2022a. doi: 10.52202/068431-2430.

Sheng Wang, Zihao Zhao, Zixu Zhuang, Xi Ouyang, Lichi Zhang, Zheren Li, Chong Ma, Tianming Liu, Dinggang Shen, and Qian Wang. Learning better contrastive view from radiologist's gaze. *Pattern Recognition*, 162:111350, 2025. doi: 10.1016/j.patcog.2025.111350.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b. doi: 10.18653/v1/2022.emnlp-main.256.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21372–21383, 2023.

Hao Yang, Hong-Yu Zhou, Jiarun Liu, Weijian Huang, Cheng Li, Zhihuan Li, Yuanxu Gao, Qiegen Liu, Yong Liang, Qi Yang, Song Wu, Tao Tan, Hairong Zheng, Kang Zhang, and Shanshan Wang. A multimodal vision–language model for generalizable annotation-free pathology localization. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-025-01574-7.

Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. Mrm: Masked relation modeling for medical image pre-training with genetics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21452–21462, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25, 2022.

Zihao Zhao, Sheng Wang, Qian Wang, and Dinggang Shen. Mining gaze for contrastive learning toward computer-assisted diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7543–7551, 2024. doi: 10.1609/aaai.v38i7.28586.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *International Conference on Learning Representations*, 2023.

A Pre-training Dataset Composition and Provenance

The pre-training corpus contains 302,047 images from 23 breast imaging datasets. Dataset-level image counts correspond to the frozen pre-training composition used in this study rather than the full sizes of the original source releases. The datasets cover screening, diagnostic, treatment, and research cohorts across mammogram, breast ultrasound, and breast MRI.

Table 4: Composition and provenance of the 23 datasets included in the pre-training corpus. NR indicates that the corresponding acquisition origin or setting could not be established from the available source documentation.

Dataset	Images	Country / Institution or Program	Population / Acquisition Setting
mammogram			
EMBED	100,217	USA / Emory Healthcare	Hospital screening and diagnostic mammogram
RSNA	36,185	USA / Emory Healthcare; Australia / BreastScreen Victoria	Screening mammogram collected at two clinical sites
Screening-DBT	13,218	USA / Duke Health System	Screening digital breast tomosynthesis studies
VinDr-Mammo	11,558	Vietnam / Hanoi Medical University Hospital and Hospital 108	Retrospective hospital mammogram examinations
CSAW-CC	6,962	Sweden / Karolinska University Hospital, Stockholm	Breast-screening case-control collection
CBIS-DDSM	2,254	NR / curated DDSM program	Curated mammographic abnormality research collection
CDD-CESM	1,354	Egypt / National Cancer Institute, Cairo University	Clinical contrast-enhanced mammogram with imaging findings and pathology
HiSBreast	714	NR	NR
DMID	396	India / Samved Hospital, Ahmedabad	Hospital digital mammogram for breast diagnosis research
MIAS	304	NR / Mammographic Image Analysis Society	Digitized mammogram research collection
INbreast	293	Portugal / Breast Centre, Centro Hospitalar de Sao Joao, Porto	University-hospital full-field digital mammogram cases
BMCD	234	Cyprus / Nicosia and Limassol General Hospitals and screening programs	Retrospective two-round sequential mammogram cases

Dataset	Images	Country / Institution or Program	Population / Acquisition Setting
CSAW-S	174	Sweden / CSAW and Karolinska-linked research program	Annotated mammogram screening cases
Breast Ultrasound			
Breast-USV	64,739	NR / Breast-USV cohort	Breast-ultrasound-video diagnosis cohort
BUS-UCLM	644	Spain / Ciudad Real General University Hospital	Hospital breast-ultrasound lesion and normal-tissue scans
BUSI	616	Egypt / Baheya Hospital, Cairo	Hospital breast ultrasound with normal, benign, and malignant categories
BrEaST-Lesions-USG	225	Poland / two oncology centers	Clinical breast-lesion ultrasound with follow-up or biopsy confirmation
OASBUD	165	Poland / Institute of Oncology, Warsaw	Benign and malignant breast-lesion ultrasound acquisitions
Breast MRI			
Duke-Breast-Cancer-MRI	58,711	USA / Duke Hospital	Retrospective biopsy-confirmed invasive breast cancer with preoperative MRI
I-SPY1	3,014	USA / multicenter ACRIN 6657 and CALGB 150007 trial	Stage II/III breast cancer during neoadjuvant chemotherapy
NACT	39	NR / Breast-MRI-NACT-Pilot, TCIA	Serial DCE-MRI during neoadjuvant chemotherapy
TCGA-BRCA-MRI	25	NR / TCGA-BRCA imaging-source-site program	Breast invasive carcinoma imaging linked to the TCGA research cohort
RIDER-Breast-MRI	6	NR / NCI RIDER program	Repeat or interval MRI for early neoadjuvant treatment response

B DiffEye Adaptation and Spatial Prior Validation

B.1 Modality-Specific DiffEye Adaptation

Because mammogram, breast ultrasound, and breast MRI differ substantially in image structure and visual-search characteristics, DiffEye is adapted independently for the three modalities. The adaptation data consist of expert eye-tracking recordings collected in our laboratory while breast images were being interpreted. For each modality, 150 images are used for model adaptation and an additional 15 held-out images are used for validation.

The objective of this modality-specific adaptation is to make the generated trajectories approximate expert visual-search behavior in the corresponding breast imaging domain. After adaptation, DiffEye is applied to the larger pre-training image corpus to generate multiple synthetic visual-search trajectories per image. These generated trajectories are subsequently aggregated into continuous attention distributions and converted into DiffEye-generated weak spatial attention priors. Thus, expert eye tracking is used to adapt and validate the trajectory generator, whereas Stage 1 consumes the generated spatial prior rather than direct expert-gaze annotations.

B.2 Trajectory-to-Spatial-Prior Construction

Multiple visual-search trajectories generated for each image are first projected into a unified image coordinate system. Spatial accumulation and smoothing produce a continuous attention heatmap, which is subsequently mapped to the visual-token space according to the image preprocessing and patch geometry. Each valid patch is thereby assigned a corresponding diagnostic attention weight.

The patch-level attention weights participate in diagnostic-evidence allocation during pre-training. Regions with higher attention are distributed between masked reconstruction and visible semantic learning. Masked diagnostic regions provide reconstruction targets, while visible diagnostic regions contribute to image–report semantic learning. The resulting masked and visible diagnostic coverage further coordinates the optimization strength of the two learning branches.

B.3 DiffEye Trajectory and Spatial-Attention Validation

The quality of DiffEye-generated spatial attention is evaluated against held-out expert gaze at both the visual-search trajectory level and the continuous attention-distribution level. Trajectory-level evaluation compares the original DiffEye and the modality-specific adapted DiffEye with corresponding expert visual-search patterns using LD, DFD, DTW, and TDE. Spatial-attention evaluation

measures the correspondence between generated attention maps and expert fixation distributions using AUC-Judd, AUC-Borji, NSS, SIM, CC, and KL divergence.

Evaluation is performed separately for mammogram, breast ultrasound, and breast MRI. The Expert–Expert rows provide human reference levels for the variability of expert visual attention, while the Original and Adapted DiffEye rows quantify the effect of modality-specific adaptation.

Table 5: Comparison of original and modality-specific adapted DiffEye with expert visual attention across mammogram, breast ultrasound, and breast MRI. The Expert–Expert rows provide human expert-reference levels. Trajectory metrics include LD, DFD, DTW, and TDE, while spatial-attention metrics include AUC-Judd, AUC-Borji, NSS, SIM, CC, and KL divergence.

			Trajectory similarity				Spatial-attention similarity					
Modality	Method	Reference	LD ↓	DFD ↓	DTW ↓	TDE ↓	AUC-Judd ↑	AUC-Borji ↑	NSS ↑	SIM ↑	CC ↑	KL ↓
mammogram	Original DiffEye	Expert gaze	17.4	4.62	2.32	116.8	0.694	0.641	0.98	0.318	0.354	2.54
	Adapted DiffEye	Expert gaze	13.6	3.55	1.69	82.7	0.836	0.778	2.05	0.558	0.646	1.10
	Expert gaze	Expert gaze reference	10.8	2.86	1.34	67.5	0.889	0.835	2.72	0.702	0.802	0.62
Ultrasound	Original DiffEye	Expert gaze	18.8	5.05	2.61	124.5	0.672	0.621	0.84	0.287	0.318	2.77
	Adapted DiffEye	Expert gaze	14.5	3.82	1.85	90.6	0.812	0.754	1.79	0.501	0.579	1.30
	Expert gaze	Expert gaze reference	12.1	3.18	1.47	74.6	0.868	0.812	2.39	0.653	0.748	0.76
MRI	Original DiffEye	Expert gaze	18.1	4.88	2.49	121.2	0.681	0.629	0.90	0.301	0.336	2.63
	Adapted DiffEye	Expert gaze	14.1	3.70	1.78	87.4	0.821	0.764	1.89	0.523	0.603	1.21
	Expert gaze	Expert gaze reference	11.6	3.05	1.42	71.2	0.878	0.822	2.51	0.674	0.771	0.69

Representative comparisons across the three breast imaging modalities are shown in Fig. 4.

C Clinical Graph Construction and Quality Control

C.1 Report-Derived Graph Construction

Clinical Graphs are constructed only after the corresponding clinical text has been accepted as an Effective Report. The graph is defined at the case level and is intended to organize clinically meaningful report semantics rather than to serve as an independent medical knowledge base. Its schema contains upper-level concepts shared across breast imaging modalities together with modality-specific concepts for mammogram, breast ultrasound, and breast MRI. Case-specific findings, imaging attributes, assessment information, pathology-related information, and their clinically meaningful relations are represented within this schema.

Graph construction uses the Effective Report as the primary semantic input. Available structured clinical facts are retained as auxiliary anchors for coverage and conflict auditing rather than being used to bypass the report. Each extracted graph element is assigned one of three provenance types. `official_fact` denotes information grounded in verified clinical facts and represented in the accepted report; `report_explicit` denotes findings or attributes explicitly described in the report; and `model_inference` denotes schema-constrained relations inferred from the report context and already extracted graph elements. Inferred elements retain confidence and inference provenance and are not allowed to overwrite verified clinical facts.

C.2 Graph Schema and Semantic Relations

Graph nodes represent clinically normalized concepts and their case-specific states. These include shared diagnostic concepts and modality-dependent attributes such as mammographic findings, ultrasound lesion characteristics, and MRI enhancement or sequence-related characteristics when supported by the corresponding report. The graph also represents assessment, pathological correlation, and other clinically supported case-level information.

Relations connect these concepts through a constrained relation schema, including hierarchical, attribute, spatial, diagnostic-support, pathology-consistency, shared-concept, and risk-order relations. The graph therefore captures both within-case semantic structure and clinically relevant relationships that can be compared across cases. Missing information is kept distinct from explicit negative

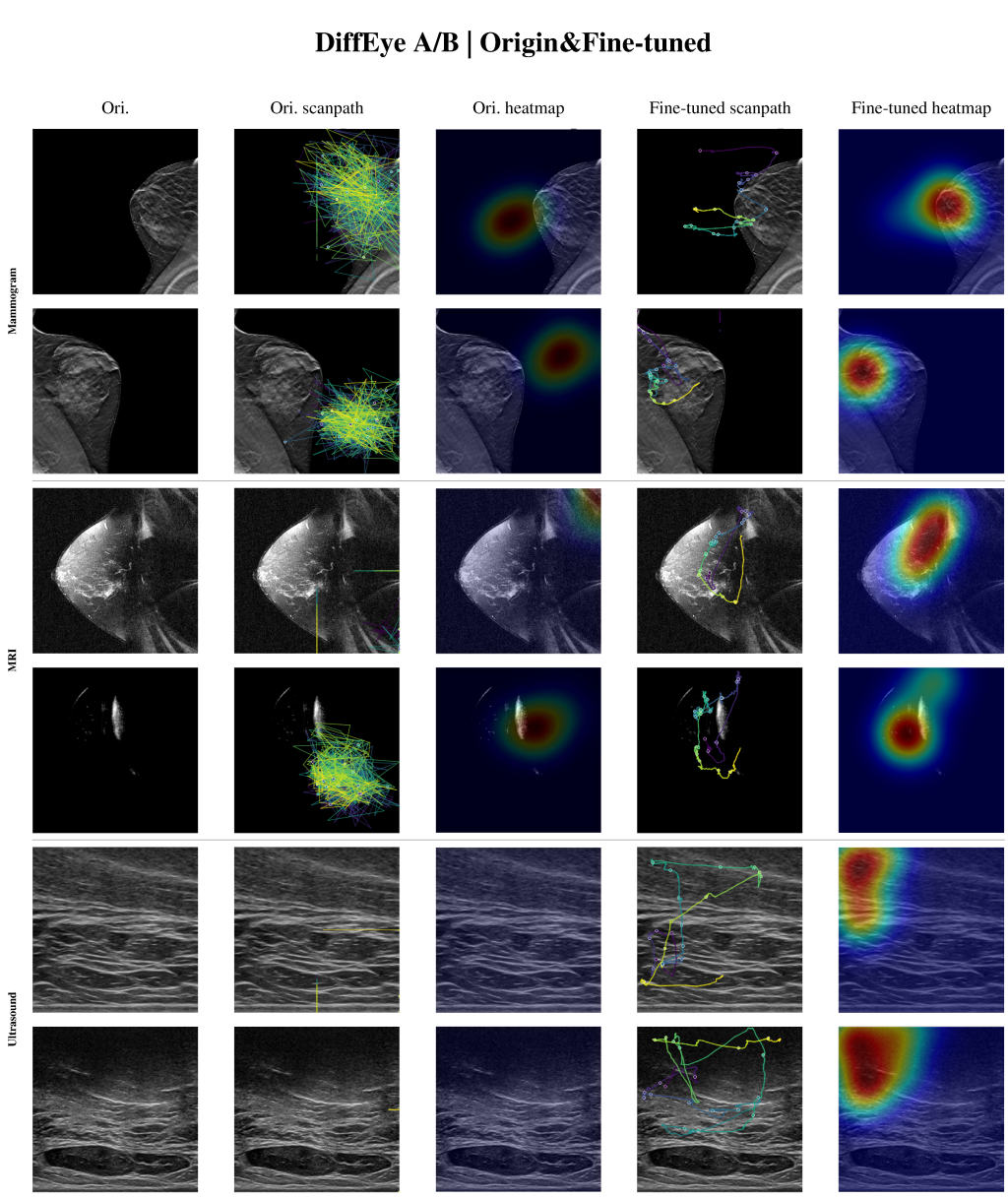


Figure 4: Comparison of DiffEye-generated visual-search trajectories and spatial-attention distributions across mammogram, breast ultrasound, and breast MRI. The visualization includes the original DiffEye, modality-specific adapted DiffEye, and corresponding expert visual-attention references.

evidence; in particular, information not mentioned in a report is not converted into an explicit absence.

C.3 Quality Control and Stage-1 Supervision

Before a Clinical Graph is admitted to Stage 1, it is checked for coverage of available verified clinical facts, schema validity, unresolved conflicts, provenance completeness, and consistency with the accepted Effective Report. If a verified fact is omitted, constrained extraction or deterministic graph repair is performed using information already supported by the Effective Report. Graphs with

unresolved conflicts or incomplete required information are not admitted as formal Stage 1 semantic assets.

The resulting graph is used as structured case-level semantic supervision. It contributes sparse inter-case semantic relations for $\mathcal{L}_{\mathrm{soft}}$, modality-aware clinical concept targets and concept prototypes for $\mathcal{L}_{\mathrm{concept}}$ and $\mathcal{L}_{\mathrm{cc}}$, and valid structured relations for $\mathcal{L}_{\mathrm{graph}}$. Clinical Graph supervision complements the image–report objectives and does not introduce direct spatial alignment between image regions and graph nodes.

D Additional Downstream Results

This section reports the downstream adaptation protocol together with supplementary analyses of image–text retrieval, clinical attribute prediction, label efficiency and external generalization, and lesion-level spatial transfer.

D.1 Downstream Evaluation Protocol

All downstream evaluations start from the final Stage 1 checkpoint selected independently of downstream test performance. Classification uses frozen image encoders and projection layers with a single trainable linear classifier. The classifier is optimized on the training partition, while checkpoint and decision-threshold selection are performed using validation data only.

For image–text retrieval, the Stage 1 image encoder, text encoder, and their projection layers remain frozen, and no retrieval-specific head or adapter is trained. Retrieval is performed directly in the normalized joint embedding space using cosine similarity.

Spatial and clinical-attribute evaluations use task-specific prediction heads. For BUSI-WHU spatial evaluation, frozen spatial representations are coupled to a 1×1 convolutional prediction head followed by bilinear upsampling. For MammosighTR clinical-attribute prediction, frozen image representations are coupled to independent categorical linear heads for BI-RADS assessment and breast composition. Task-specific configurations are fixed before test evaluation.

All trainable downstream experiments use the predefined seeds 11, 12, and 13. Model selection, early stopping, and threshold selection, when applicable, are restricted to the validation partition. The test partition is used only for the final evaluation.

D.2 Image–Text Retrieval

Bidirectional image–text retrieval evaluates whether visual and clinical-text representations form a coherent joint semantic space. Image-to-text and text-to-image retrieval are evaluated for all three imaging modalities. mammogram uses exam- or breast-level image–report pairs, breast ultrasound uses lesion-case–report pairs, and breast MRI uses study–report pairs.

For GazeBreast-FM, retrieval is evaluated in a frozen zero-shot setting. The Stage 1 image encoder, text encoder, and their projection layers are kept fixed, and no retrieval-specific adapter or task head is trained. Image and text representations are projected into the shared embedding space, ℓ_2-normalized, and ranked according to their pairwise cosine similarity. The same frozen representations are used for both image-to-text and text-to-image retrieval.

Both retrieval directions are evaluated using Recall@1, Recall@5, and Recall@10 as the primary metrics. Higher recall indicates stronger correspondence between visual representations and clinical text within the shared semantic space.

Table 7: Clinical attribute prediction across three breast imaging modalities.

Modality	Dataset	Model	Macro-AUROC ↑	Macro-AUPRC ↑	Macro-F1 ↑
mammogram	MammosighTR	DINOv2 ViT-B/14	0.8411 ± 0.0009	0.6129 ± 0.0017	0.5562 ± 0.0015
		MedSigLIP-448	0.8513 ± 0.0004	0.6265 ± 0.0003	0.5578 ± 0.0001
		MAM-CLIP	0.7790 ± 0.0005	0.5144 ± 0.0002	0.4452 ± 0.0085
		Ours	$\mathbf{0.8876 \pm 0.0016}$	$\mathbf{0.6843 \pm 0.0028}$	$\mathbf{0.6195 \pm 0.0036}$
Ultrasound	BCMID v3	DINOv2 ViT-B/14	0.5029 ± 0.0625	0.2109 ± 0.0297	0.0934 ± 0.0060
		MedSigLIP-448	0.6696 ± 0.0545	0.4173 ± 0.0166	0.0948 ± 0.0073
		OpenUS	0.6472 ± 0.0319	0.2800 ± 0.0204	0.0935 ± 0.0006
		Ours	$\mathbf{0.7148 \pm 0.0306}$	$\mathbf{0.4598 \pm 0.0148}$	$\mathbf{0.1105 \pm 0.0054}$
MRI	FGT 14921217	DINOv2 ViT-B/14	0.7985 ± 0.0140	0.5150 ± 0.0062	0.4047 ± 0.0209
		MedSigLIP-448	0.8390 ± 0.0025	0.5882 ± 0.0100	0.4260 ± 0.0208
		Pillar-0 BreastMRI	0.8527 ± 0.0047	0.6074 ± 0.0092	0.4418 ± 0.0170
		Ours	$\mathbf{0.8786 \pm 0.0042}$	$\mathbf{0.6421 \pm 0.0086}$	$\mathbf{0.4712 \pm 0.0146}$

Table 6: Bidirectional image–text retrieval across three breast imaging modalities. Results are reported as Recall@1, Recall@5, and Recall@10 for image-to-text (I→T) and text-to-image (T→I) retrieval.

Modality	Dataset	Model	I→T R@1↑	I→T R@5↑	I→T R@10↑	T→I R@1↑	T→I R@5↑	T→I R@10↑
mammogram	BCMID v3	OpenAI CLIP	0.3210	0.3740	0.4210	0.3180	0.3690	0.4160
		MedSigLIP-448	0.3310	0.3960	0.4530	0.3290	0.3910	0.4490
		Mammo-CLIP	0.3470	0.4320	0.5040	0.3440	0.4260	0.4960
		GazeBreast-FM	**0.3680**	**0.4760**	**0.5630**	**0.3640**	**0.4690**	**0.5540**
Ultrasound	LIJUN BREAST US REPORT	Jina-CLIP-v2	0.3027	0.3085	0.3171	0.3098	0.3100	0.3185
		BiomedCLIP	0.3014	0.3043	0.3114	0.3000	0.3057	0.3145
		SonoCLIP	0.3014	0.3114	0.3242	0.3014	0.3128	0.3284
		GazeBreast-FM	**0.3043**	**0.3199**	**0.3384**	**0.3057**	**0.3228**	**0.3441**
MRI	BREAST-DIAGNOSIS	OpenAI CLIP	0.3000	0.2632	0.5263	0.1053	0.3158	0.5789
		MedSigLIP-448	0.3526	0.5632	0.6316	0.3526	0.3158	0.5263
		Pillar-0 BreastMRI	0.4053	0.3684	0.6842	0.4053	0.3684	0.6316
		GazeBreast-FM	**0.4579**	**0.5737**	**0.7895**	**0.4579**	**0.4211**	**0.7895**

D.3 Clinical Attribute Prediction

Clinical attribute prediction evaluates whether the learned representations encode fine-grained, modality-specific clinical semantics beyond disease-level classification. mammogram is evaluated on MammosighTR using BI-RADS assessment and Breast Composition/Density, with the two attribute tasks aggregated for overall evaluation. Breast ultrasound is evaluated on BCMID v3 using seven-class BI-RADS category prediction (0–6). Breast MRI is evaluated on FGT 14921217 using four-class Fibroglandular Tissue prediction (A/B/C/D). Across the three modalities, we report Macro-AUROC, Macro-AUPRC, and Macro-F1 for a consistent comparison of clinical attribute representations.

D.4 Label Efficiency and External Generalization

Label efficiency is evaluated on BUSI-WHU using 10%, 50%, and 100% of the available training annotations. GazeBreast-FM achieves AUROC values of 0.7846 ± 0.0078, 0.8593 ± 0.0049, and 0.8826 ± 0.0037 under the three supervision levels, respectively.

External generalization is evaluated by transferring the BUSI-WHU classification models to the independent QAMEBI dataset. GazeBreast-FM achieves an AUROC of 0.8818 ± 0.0046, corresponding to an AUROC change of -0.0008 relative to the fully supervised BUSI-WHU result.

Table 8: Label efficiency on BUSI-WHU and external classification generalization to QAMEBI.

Model	BUSI 10%	BUSI 50%	BUSI 100%	QAMEBI AUROC
OpenUS	0.6789 ± 0.0227	0.8237 ± 0.0043	0.8380 ± 0.0047	0.7125 ± 0.0098
DINOv2 ViT-B/14	0.6711 ± 0.0245	0.7749 ± 0.0099	0.8134 ± 0.0053	0.6834 ± 0.0087
MedSigLIP-448	0.7269 ± 0.0006	0.8038 ± 0.0016	0.8178 ± 0.0011	0.8163 ± 0.0009
Ours	$\mathbf{0.7846 \pm 0.0078}$	$\mathbf{0.8593 \pm 0.0049}$	$\mathbf{0.8826 \pm 0.0037}$	$\mathbf{0.8818 \pm 0.0046}$

D.5 Additional Spatial Evaluation

QAMEBI provides an additional evaluation of cross-dataset lesion-level spatial transfer in breast ultrasound. Segmentation performance is measured using Dice, and lesion localization is evaluated using mAP@0.5.

GazeBreast-FM achieves a Dice score of 0.8624 ± 0.0038 and an mAP@0.5 of 0.8781 ± 0.0049 on QAMEBI, demonstrating that the learned lesion-level representation remains transferable across independent ultrasound datasets.

Table 9: Additional lesion segmentation and localization evaluation on QAMEBI.

Model	Dice	mAP@0.5
DINOv2 ViT-B/14	0.7026 ± 0.0013	0.6679 ± 0.0111
MedSigLIP-448	0.7402 ± 0.0005	0.6897 ± 0.0021
SegFormer-B2	0.8106 ± 0.0009	0.8389 ± 0.0060
Ours	$\mathbf{0.8624 \pm 0.0038}$	$\mathbf{0.8781 \pm 0.0049}$