A SEMANTICALLY ALIGNED AND ENRICHED VISION-LANGUAGE MODEL FOR BREAST CANCER DIAGNOSIS ON ROUTINE CHEST CT

Anonymous authors
Paper under double-blind review

ABSTRACT

Routine chest CT contains the breast tissue, however, physicians often omit breast findings due to their low contrast in CT scans. In this work, we explore whether the weakly-supervised vision-language contrastive learning provides scalable opportunistic breast cancer diagnosis using routine chest CT and reports. First, a benchmark dataset is constructed (BreastCT-12K), which comprises 12,696 patients with routine chest CT scans, CT reports, and temporally matched breast MRI reports. Since CT reports often lack breast description, we curate MRI reports to retain CT-visible findings to enrich the paired CT reports. We then develop a Report-Enriched and Aligned CHest CT vision-language model for Breast cancer diagnosis (**REACH-breast**). Since whole-volume contrastive pretraining may overlook small and side-specific abnormalities of breast, REACH-breast introduces two modules to enhance fine-grained feature alignment and representation: 1) Breast-Centric Hierarchical Alignment (BCHA) matches the left, right, and bilateral breast representations with corresponding text descriptions and bridges the bilateral and global context, preserving laterality while making whole-volume features breast-aware. 2) Side-aware Evidence–Patch Association (SEPA) links report evidence to specific breast-region within patches, strengthening fine-grained visual–semantic correspondence without lesion-level annotations. Using chest CT and fixed class prompts at inference, REACH-breast achieves 89.19% AUC for zero-shot malignancy classification on the test split of BreastCT-12K, significantly exceeding the state-of-the-art breast-adapted zero-shot foundation models by at least 7.08%. It also yields the best performance in bidirectional CT-report retrieval and report generation tasks. External evaluation confirms superior performance of REACH-breast. These results demonstrate the effectiveness of vision-language modeling for opportunistic breast assessment on routine chest CT.

1 INTRODUCTION

Breast cancer remains a major cause of cancer mortality in women, making early detection an important goal (Bray et al., 2024). Although dedicated breast imaging is central to screening and diagnosis, routine chest CT also captures breast tissue and can reveal clinically relevant abnormalities (Park et al., 2023; Thomas et al., 2025). Recent progress in opportunistic CT screening highlights the value of examining anatomy beyond the original clinical indication (Cao et al., 2023; Hu et al., 2025; Chen et al., 2026). This motivates learning breast-aware CT representations that can support a range of downstream tasks.

CLIP-style vision–language pretraining (VLP) provides a natural approach to learning such transferable representations through image–text alignment (Radford et al., 2021). CT-CLIP, fVLM, and ViSD-Boost extend this paradigm to 3D CT, enabling zero-shot recognition and downstream transfer (Hamamci et al., 2024; Shui et al., 2025; Cao et al., 2025). However, report-supervised CT VLP depends on what radiologists document. Routine chest CT reports prioritize intrathoracic findings, and incidental breast abnormalities often remain unreported (Park et al., 2023). As illustrated in Figure 1(a,b), breast tissue is present in the image but underrepresented in the paired report, leaving models with insufficient supervision for breast-specific visual–semantic associations. Moreover, whole-volume alignment provides limited guidance for localized, side-specific breast findings.

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.

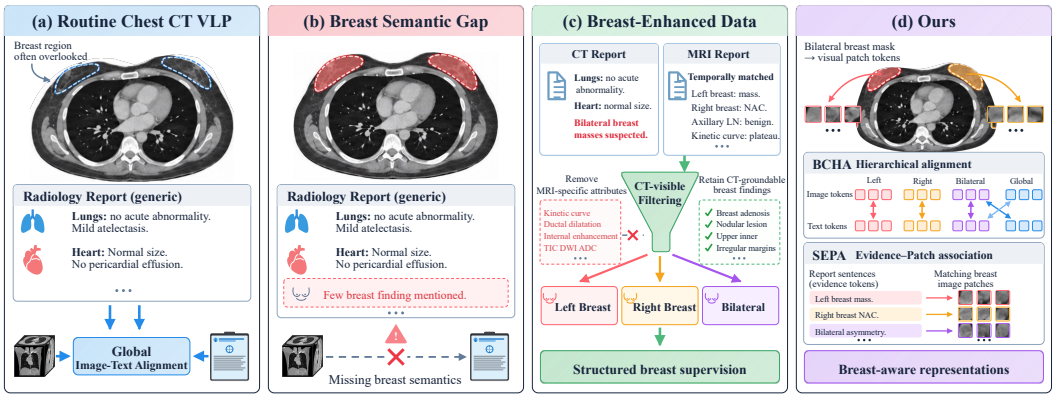


Figure 1: (a) Routine CT VLP, (b) breast findings missing breast semantics, (c) MRI-derived breast supervision, and (d) our breast-aware semantic alignment.

These limitations motivate two complementary requirements: enriching breast-related language supervision and grounding that supervision at the appropriate anatomical and spatial scales.

We introduce **REACH-breast**, to our knowledge the first breast-enhanced CT vision–language model for transferable 3D breast representation learning. We first pretrain a general 3D CT vision–language model as our baseline on approximately 200,000 chest CT–report pairs. We then perform breast-enhanced training on the training split of **BreastCT-12K**, our newly constructed dataset of 12,696 patients with routine chest CT, corresponding CT reports, and temporally matched breast MRI reports. We supplement the CT reports with CT-groundable breast descriptions curated from the MRI reports and construct left-breast, right-breast, and bilateral texts for global and regional supervision (Figure 1(c)).

To connect this supervision with the appropriate anatomical regions, we introduce **Breast-Centric Hierarchical Alignment (BCHA)**. Breast masks guide the pooling of left, right, and bilateral breast representations from a shared set of CT patch tokens. BCHA aligns each regional representation with its corresponding description and adds cross-level alignment between global and bilateral views. This links breast-specific semantics to whole-volume representations while preserving side-specific supervision.

Regional alignment alone does not establish which local visual patterns support individual findings. We therefore introduce **Side-aware Evidence–Patch Association (SEPA)**, which associates sentence-level clinical evidence with CT patch tokens within the corresponding breast region. Evidence–patch similarities are aggregated through max-similarity matching and optimized with a patient-level contrastive objective. This encourages fine-grained visual–semantic correspondence without requiring lesion-level annotations.

Experiments demonstrate the effectiveness and transferability of REACH-breast across a range of breast CT tasks. Our model achieves the highest zero-shot diagnostic AUC among the compared methods on both internal and external test sets (89.19% and 83.73%, respectively). It substantially improves bidirectional image–text retrieval and downstream performance in report generation and BI-RADS classification. Ablation studies support the benefits of MRI-derived supervision, hierarchical alignment, and local evidence association. Our key contributions are summarized as follows:

- We construct BreastCT-12K, a breast-enhanced CT–language dataset of 12,696 patients, leveraging temporally matched MRI reports to supplement sparse breast descriptions in CT reports with CT-groundable, side-specific and bilateral textual supervision.
- We introduce REACH-breast, a framework composed of two modules, BCHA and SEPA, that establishes anatomical-view-level alignment and fine-grained evidence–patch association without requiring lesion-level annotations.
- We demonstrate the effectiveness and transferability of the learned representations through zero-shot diagnosis, report generation, bidirectional image–text retrieval, and BI-RADS classification, complemented by cross-center evaluation, component-wise ablations, and qualitative visualizations.

2 Related Work

2.1 Medical Vision-Language Pre-training

Medical vision–language pretraining (VLP) learns transferable representations from images and reports. ConVIRT and BioViL use image–report contrastive learning (Zhang et al., 2022; Boecking et al., 2022); GLoRIA and MGCA add region, word, or disease-level correspondence (Huang et al., 2021; Wang et al., 2022a). MedKLIP and ASG ground clinical entities or sentences in anatomy (Wu et al., 2023; Li et al., 2024), while MedCLIP uses semantic similarity to learn from unpaired data Wang et al. (2022b). For volumetric imaging, BIUD distills chest X-ray knowledge into CT Cao et al. (2024), and CT-CLIP aligns volumes with reports Hamamci et al. (2024). CT-GLIP and fVLM introduce anatomical alignment (Lin et al., 2024; Shui et al., 2025); ViSD-Boost and Disease-Centric VLP further emphasize abnormal or disease-level signals (Cao et al., 2025; Shi et al., 2026). These methods strengthen general CT representation learning, but breast findings remain sparse in routine chest CT reports. Unlike CT-CLIP's volume–report alignment and the anatomy- or disease-focused objectives of fVLM and ViSD-Boost, REACH-breast first adds CT-groundable breast descriptions curated from paired MRI reports. It then aligns global, left, right, and bilateral breast views and associates individual evidence segments with breast patches. The distinction is therefore both richer breast-specific language supervision and alignment at the relevant anatomical scales, without lesion-level annotations.

2.2 Anatomy-Aware Alignment and Weakly Supervised Grounding

Local correspondence extends medical vision–language learning beyond global image–report matching. GLoRIA links report words to attended image regions, while LoVT and MGCA learn local associations at multiple scales (Huang et al., 2021; Müller et al., 2022; Wang et al., 2022a). MedKLIP aligns structured clinical entities with patches; AGXNet combines observation–anatomy relations and positive–unlabeled learning, and ASG uses anatomical region–sentence pairs (Wu et al., 2023; Yu et al., 2022; Li et al., 2024). These approaches show how reports can guide spatial learning without lesion-level annotations. For volumetric CT, Ichinose et al. (2023) combine anatomical segmentation with report structuring but use annotated region–description pairs. CT-GLIP constructs organ-level CT–text pairs, whereas fVLM matches anatomical regions with reports using disease-aware pairing (Lin et al., 2024; Shui et al., 2025). In mammography, MaMA exploits multi-view and multi-scale correspondence, while GLAM adds geometry-guided local alignment (Du et al., 2025b;a). Routine chest CT poses a different challenge: breast findings are small, side-specific, and sparsely documented. We enrich CT reports with breast descriptions, couple left, right, and bilateral representations to global context through BCHA, and associate sentence evidence with local 3D patches through SEPA, without lesion-level annotations.

3 Method

3.1 Framework Overview

REACH-breast learns breast-aware CT representations from enriched text and alignment at anatomical and local scales (Figure 2). Its 3D vision transformer and radiology-domain BERT are initialized from a general CT vision–language baseline pretrained on approximately 200,000 chest CT–report pairs from the CT-RATE training split and an in-house corpus.

Breast-enhanced training uses BreastCT-12K. Filtered MRI reports enrich the paired CT reports and provide left, right, and bilateral textual views. One CT encoder pass yields shared patch tokens for global and breast-region pooling. Beyond the baseline global CT–report alignment, Breast-Centric Hierarchical Alignment (BCHA) aligns breast regions with their texts and couples global and bilateral representations. Side-aware Evidence–Patch Association (SEPA) further relates sentence-level report evidence to encoded breast patches.

3.2 Breast-enhanced Dataset Construction

BreastCT-12K contains 12,696 patients with a routine chest CT, its report, and a breast MRI report acquired within 90 days. We apply the same MRI-report filtering and text-construction procedure

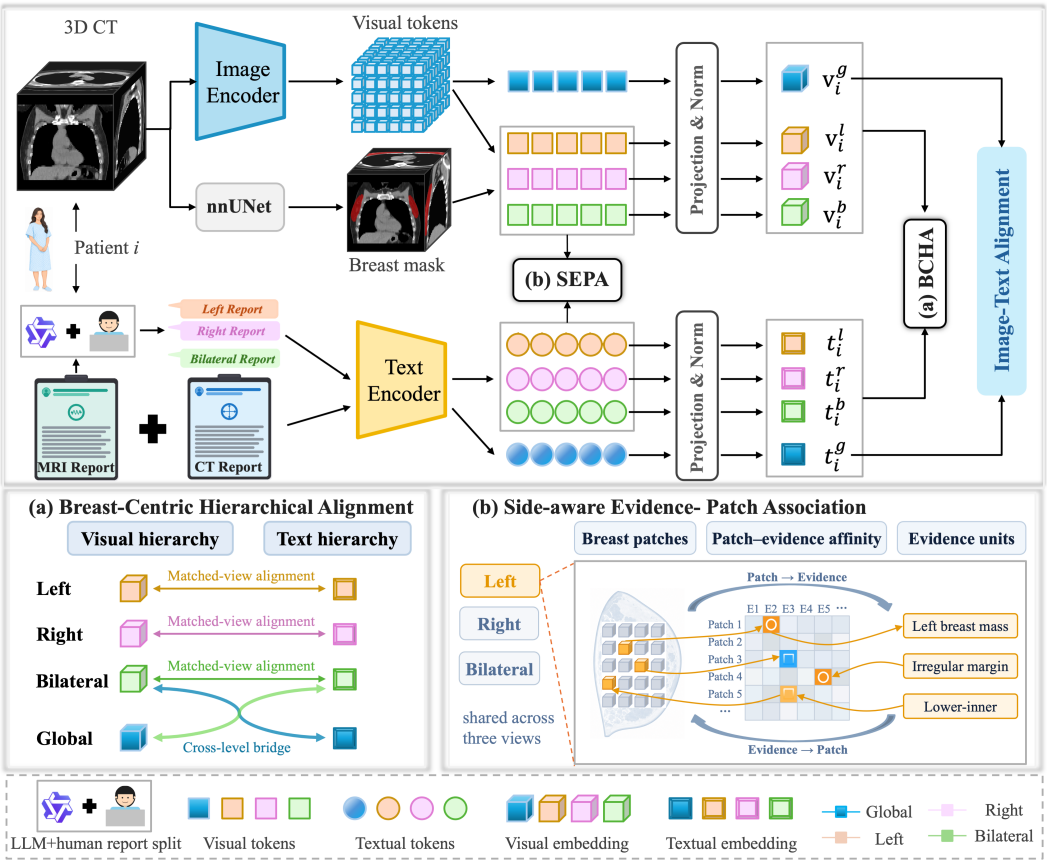


Figure 2: Overview of REACH-breast. (a) BCHA aligns breast views and bridges bilateral and global features; (b) SEPA associates sentence evidence with breast patches.

to both the training and internal test splits, producing enhanced reports and breast-specific texts for each split. A Large Language Model (LLM) first screens MRI statements for potentially CT-observable content; human review then checks the selected descriptions against the source report (detailed in Appendix A.2). We keep information assessable on CT, such as the presence of a mass, its side and location, approximate size, and margins. We remove MRI-specific details that CT cannot depict, including ductal details, signal characteristics, and enhancement patterns. BI-RADS assessment terms are retained in the curated *Impression* as report context, but they do not specify lesion coordinates on CT.

We first construct the enhanced global text by combining the original CT report with the filtered MRI breast descriptions. We then partition the filtered MRI report into left-breast, right-breast, and bilateral texts. Left and right texts describe their respective sides, while bilateral text captures shared or comparative findings. These distinct views prevent a unilateral statement from supervising the opposite breast; a missing regional description is omitted from its loss rather than treated as a normal finding. An nnU-Net pipeline supplies left, right, and bilateral breast masks (Isensee et al., 2021) that delimit breast regions, not lesions. Thus patient i has CT x_i, global text r_i^g, and regional text–mask pairs $\{r_i^s, m_i^s\}_{s\in\mathcal{S}}$ for $\mathcal{S} = \{l, r, b\}$, as shown in Figure 2.

3.3 Breast-Centric Hierarchical Alignment

Global CT–enhanced-report alignment is retained as the separate baseline loss $\mathcal{L}_{\text{global}}$. It supervises whole-volume semantics but does not indicate laterality or bilateral relations. BCHA adds two complementary forms of supervision (Figure 2a): matched left, right, and bilateral breast pairs, followed by cross-level pairings between bilateral and global views.

One full-volume CT pass produces contextualized patch tokens $P_i = E_v(x_i) \in \mathbb{R}^{N \times d_v}$. For each view $s \in \mathcal{S} = \{l, r, b\}$, we retain patches Ω_i^s whose mean breast-mask coverage is at least $\rho = 0.1$; the bilateral mask spans both breasts. Global attention pooling A_g uses all tokens, while regional pooling A_m uses the selected patches. The latter is initialized from A_g and shared across views with a learned view query. Since selection follows full-volume self-attention, breast patches retain surrounding CT context. Projection ϕ_v and ℓ_2 normalization yield

$$v_i^g = \mathrm{Norm}(\phi_v(A_g(P_i))), \qquad v_i^s = \mathrm{Norm}(\phi_v(A_m(P_i, \Omega_i^s; s))), \qquad t_i^u = \mathrm{Norm}(E_t(r_i^u)), \quad (1)$$

where $s \in \mathcal{S}$ and $u \in \{g, l, r, b\}$. The bilateral feature is pooled directly from both breast regions; all views share a contrastive space.

Anatomically matched alignment. A left-sided mass supervises the left breast, and a right-sided finding supervises the right. These pairs encourage laterality-aware features rather than matches to the opposite breast or unrelated thoracic anatomy. The bilateral pair supplies joint and comparative cues, including density differences and departures from symmetry. Each pair uses symmetric contrastive loss $\ell_\tau(V, T; \mathcal{I})$, with same-patient positives and other mini-batch patients as negatives in both directions (Appendix A.5). The baseline term is $\mathcal{L}_{\text{global}} = \ell_\tau(V^g, T^g; \mathcal{I}_g)$ for all patients. For regional view s, let $\mathcal{I}_s$ include patients with valid mask and text and set $\mathcal{L}_s = \ell_\tau(V^s, T^s; \mathcal{I}_s)$. Then

$$\mathcal{L}_{\text{side}} = \frac{1}{|\mathcal{U}|} \sum_{s \in \mathcal{U}} \mathcal{L}_s, \qquad \mathcal{L}_{\text{bilateral}} = \mathcal{L}_b, \quad \mathcal{U} = \{s \in \{l, r\} : |\mathcal{I}_s| > 0\}. \quad (2)$$

An undescribed side is omitted, not treated as negative; terms without eligible patients are zero.

Cross-level alignment. To carry breast semantics into the whole-volume representation, we align global CT features V^g with bilateral breast text T^b. Conversely, aligning bilateral CT features V^b with enhanced global text T^g grounds the breast region in the complete report context. This reciprocal supervision promotes cross-scale semantic agreement:

$$\mathcal{L}_{\text{cross}} = \frac{1}{2} \left[\ell_\tau(V^g, T^b; \mathcal{I}_b) + \ell_\tau(V^b, T^g; \mathcal{I}_b) \right]. \quad (3)$$

The BCHA objective is defined as the sum of the three components:

$$\mathcal{L}_{\text{BCHA}} = \mathcal{L}_{\text{side}} + \mathcal{L}_{\text{bilateral}} + \mathcal{L}_{\text{cross}}. \quad (4)$$

3.4 Side-aware Evidence–Patch Association

SEPA adds finer correspondence within each breast view (Figure 2b). A report may describe several findings in different locations. SEPA lets each sentence-level evidence unit find its closest CT patch and each selected patch find its closest evidence unit, complementing BCHA's regional alignment without lesion annotations.

For left, right, and bilateral views, SEPA uses the shared encoders' contextualized outputs rather than pooled BCHA embeddings. We split curated *Findings* and *Impression* into section-labeled segments joined by `[SEP]`. The encoded state at each separator summarizes one evidence unit; after the text pooler's dense layer, projection, and normalization, up to 12 units form $Q_i^s = \{q_{i,m}^s\}_{m=1}^{M_i^s}$. From the corresponding breast mask, we select up to 256 patches in Ω_i^s with highest coverage. Their projected, normalized encoder tokens form $Z_i^s = \{z_{i,k}^s\}_{k=1}^{K_i^s}$.

For image i and text j in view s, $S_{i,j}^s(k, m) = (z_{i,k}^s)^\top q_{j,m}^s$ measures the similarity between patch k and evidence unit m. Text-to-patch matching selects each unit's most similar patch; patch-to-text matching selects each patch's most similar unit. Averaging these best matches gives two complementary pair scores:

$$\begin{aligned} a_{i,j}^{s,t\to p} &= \frac{1}{M_j^s} \sum_{m=1}^{M_j^s} \max_{1 \le k \le K_i^s} S_{i,j}^s(k, m), \\ a_{i,j}^{s,p\to t} &= \frac{1}{K_i^s} \sum_{k=1}^{K_i^s} \max_{1 \le m \le M_j^s} S_{i,j}^s(k, m). \end{aligned} \quad (5)$$

Max matching allows different findings to select different patches: a unilateral nodule can favor its own side, while bilateral density evidence can involve both breasts.

To learn these associations, we score every eligible CT–report pair in the mini-batch. Same-patient pairs occupy the diagonal of $A^{s,t\to p}$ and $A^{s,p\to t}$; other patients are negatives. For valid patients $\mathcal{J}_s \subseteq \mathcal{I}_s$, diagonal labels y, and learned temperature τ_e, the symmetric contrastive objective is

$$\mathcal{C}_{\tau_e}(A) = \tfrac{1}{2}\big[\mathrm{CE}(A/\tau_e, y) + \mathrm{CE}(A^\top/\tau_e, y)\big]. \tag{6}$$

It rewards the matched patient in both CT-to-report and report-to-CT directions. We apply it to both matching scores and average the available breast views:

$$\mathcal{L}^s_{\mathrm{SEPA}} = \tfrac{1}{2}\big[\mathcal{C}_{\tau_e}(A^{s,t\to p}) + \mathcal{C}_{\tau_e}(A^{s,p\to t})\big], \qquad \mathcal{L}_{\mathrm{SEPA}} = \frac{1}{|\mathcal{S}_e|}\sum_{s\in\mathcal{S}_e}\mathcal{L}^s_{\mathrm{SEPA}}. \tag{7}$$

Here CE is row-wise cross-entropy, $\mathcal{S}_e = \{s \in \mathcal{S} : |\mathcal{J}_s| > 1\}$, and an empty $\mathcal{S}_e$ gives zero loss. SEPA and BCHA update the shared encoders at local and anatomical scales, respectively.

3.5 Training Objective

We optimize global alignment, BCHA, and SEPA with

$$\mathcal{L} = \lambda_g\mathcal{L}_{\mathrm{global}} + \lambda_h\mathcal{L}_{\mathrm{BCHA}} + \lambda_e\mathcal{L}_{\mathrm{SEPA}}, \tag{8}$$

where λ_g, λ_h, and λ_e balance the three loss terms and are set to 1, 1, and 0.5, respectively. SEPA is zero during the first three completed epochs and activates afterward.

4 Experiments

4.1 Experimental Setup

Datasets. Our general pretraining corpus comprises approximately 200,000 chest CT–report pairs in total from the CT-RATE training split and an in-house hospital data. The breast-specific dataset comprises BreastCT-12K and external cohorts. Across both breast cohorts, malignant status is confirmed by histopathology. Non-malignant status is established by benign histopathology or at least two years of follow-up without evidence of malignancy.

BreastCT-12K comprises 12,696 patients (3,918 malignant and 8,778 non-malignant), each with a chest CT scan, its report, and a paired breast MRI report. The CT and MRI examinations were performed within 90 days of each other. For patients with pathological confirmation, both examinations were performed within the six months preceding pathological diagnosis. An approximately 9:1 patient-level stratified split yields 11,426 training patients (3,526 malignant and 7,900 non-malignant). The held-out internal test set contains 1,270 patients (392 malignant and 878 non-malignant).

Table 1: Zero-shot diagnosis on the BreastCT-12K test set and external cohort (%). All zero-shot models are adapted on the same BreastCT-12K training split. Best and second-best results are bold and underlined, respectively. External results are unweighted averages of center-specific metrics.

	Method	Internal				External			
		AUC	ACC	F1	Prec	AUC	ACC	F1	Prec
Supervised	Base model	80.19	78.50	52.36	82.87	62.55	58.84	39.15	58.78
	CT-LiPro	80.67	76.38	60.42	62.57	51.98	44.57	15.42	46.98
	CT-VocabFine	82.39	77.64	55.06	72.50	50.75	48.21	9.41	45.55
Zero-shot	CT-CLIP	75.82	71.10	<u>54.41</u>	53.03	47.64	47.08	33.52	45.86
	Merlin	80.63	74.80	37.50	<u>80.00</u>	60.72	56.65	27.09	58.48
	fVLM	68.16	61.80	53.01	43.33	<u>68.70</u>	60.62	58.71	51.38
	ViSD-Boost	<u>82.11</u>	<u>76.32</u>	47.18	79.76	65.96	<u>76.88</u>	<u>60.22</u>	<u>67.16</u>
	Ours	**89.19**	**86.54**	**77.41**	**80.27**	**83.73**	**77.46**	**65.63**	**71.13**

The external cohort contains 5,526 patients (1,913 malignant and 3,613 non-malignant) from three centers held out from training and hyperparameter selection. AUC, accuracy, F1-score, and precision are computed per center and averaged with equal weight across centers (details in Appendix A.6).

Evaluation metrics. Following CT-CLIP (Hamamci et al., 2024), we report AUC, accuracy, sensitivity, specificity, precision, and F1 for zero-shot diagnosis. Metrics for report generation, BI-RADS risk-group classification, and bidirectional retrieval are detailed in Appendix A.1.

Implementation details. Preprocessing, model and optimization settings, and comparison protocols are detailed in Appendix A.2.

4.2 Zero-shot Diagnosis

We evaluate prompt-based malignancy classification on the BreastCT-12K internal test set and an independent external cohort. For the breast-adapted comparison, all VLP models undergo breast-specific vision–language adaptation on the same BreastCT-12K training split and use the fixed class prompts in Appendix A.3. Supervised methods provide additional references.

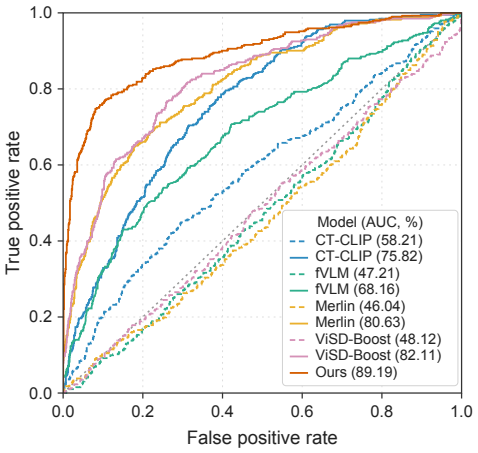


Figure 3: ROC curves before (dashed) and after (solid) training on BreastCT-12K, with REACH-breast shown in orange.

Internal and external diagnosis. As shown in Table 1, REACH-breast achieves the best results across all four internal metrics among the zero-shot methods, including 89.19% AUC and 77.41% F1. Its AUC exceeds the strongest breast-adapted comparator, ViSD-Boost, by 7.08 percentage points. On the external cohort, REACH-breast also leads all four metrics, achieving 83.73% AUC, 15.03 points above the next-best fVLM result. These results support the effectiveness of our breast-aware representations and their transferability across centers. Model-specific case availability is detailed in Appendix A.4.

Effectiveness of BreastCT-12K. To assess the utility of our dataset, we compare four CT VLP models before and after training on BreastCT-12K. As shown in Figure 3, all four models improve on the internal test set, with AUC increasing from 46.04–58.21% to 68.16–82.11%. These consistent gains across different models support the effectiveness of BreastCT-12K in providing breast-specific supervision through enriched CT–report pairs. Together with the results in Table 1, they highlight the complementary contributions of our dataset and alignment framework. Full off-the-shelf results are provided in Appendix Table 6.

4.3 Radiology Report Generation

To assess whether breast-aware representations support clinically relevant report generation, we compare frozen and fine-tuned image encoders using identical decoder architectures and training splits (Table 2). REACH-breast achieves the highest F1 in both settings (66.20% and 66.31%, respectively). Strong performance with a frozen encoder indicates that the learned representations already encode information useful for report generation. Qualitative examples in Appendix G further demonstrate that REACH-breast captures breast-specific findings omitted by the baseline, supporting more comprehensive breast-related reporting.

4.4 Image-text Retrieval

REACH-breast ranks first across all image-to-text and text-to-image retrieval metrics on the BreastCT-12K internal test set (Table 3). Its image-to-text R@1 exceeds Merlin by 3.62 percentage points, and image-to-text mAP reaches 11.49% versus 6.88%. The gain also holds in the reverse direction, with a text-to-image mAP of 10.26% versus 7.09%. These bidirectional improvements support stronger alignment between individual CT volumes and their enriched descriptions.

Table 2: Report generation on BreastCT-12K (%). Models use the same training split and decoder architecture; P/R: precision/recall. Best/second-best values are bold/underlined, including ties.

Setting	Model	ACC	P	R	F1	BLEU-4	METEOR	ROUGE-L
Frozen	CT-CLIP	58.09	64.84	67.21	64.50	36.24	56.66	52.01
	Merlin	58.95	66.69	65.74	64.92	**36.52**	53.94	51.17
	fVLM	51.22	59.90	60.46	58.17	20.68	47.08	43.53
	ViSD-Boost	55.86	66.86	62.38	63.34	35.05	54.13	53.42
	Ours	**59.96**	**67.21**	**67.67**	**66.20**	36.30	**57.08**	**54.08**
Fine-tuning	CT-CLIP	58.59	65.12	67.42	64.81	36.63	56.77	52.22
	Merlin	**60.22**	67.23	67.50	66.13	38.03	55.93	53.15
	fVLM	55.43	65.14	64.99	62.85	34.41	55.26	52.62
	ViSD-Boost	56.00	67.40	63.32	63.28	35.03	54.60	53.24
	Ours	**60.22**	**67.44**	**67.83**	**66.31**	**38.29**	**57.38**	**54.01**

Table 3: Bidirectional retrieval between CT volumes and breast-enriched reports on the BreastCT-12K test set (%). Best and second-best results are bold and underlined, respectively.

Model	Image-to-text retrieval					Text-to-image retrieval				
	R@1	R@5	R@10	R@50	mAP	R@1	R@5	R@10	R@50	mAP
CT-CLIP	0.24	1.18	3.39	16.77	1.82	0.47	2.44	4.65	20.00	2.39
Merlin	2.44	10.16	14.25	35.20	6.88	2.36	10.32	14.80	36.77	7.09
fVLM	0.24	0.87	1.58	8.29	1.13	0.16	1.03	2.05	10.02	1.20
ViSD-Boost	0.32	1.18	3.24	12.39	1.60	0.24	1.74	4.03	13.89	1.80
Ours	**6.06**	**15.75**	**21.42**	**40.47**	**11.49**	**4.09**	**18.50**	**20.47**	**42.06**	**10.26**

4.5 BI-RADS Classification

We evaluate BI-RADS classification across low-, intermediate-, and high-risk groups, with grouping criteria detailed in Appendix A.1. As shown in Figure 4, REACH-breast leads all four metrics, achieving 78.74% accuracy and 88.37% macro one-vs-rest AUC. Compared with Merlin, it improves macro recall by 11.02%. These results highlight its potential for CT-based breast cancer risk stratification and clinical management. The complete results are in Appendix Table 7.

4.6 Ablation Study

We conduct an ablation study on BreastCT-12K to assess the contributions of report enrichment, BCHA, and SEPA (Table 4). Starting from the general CT baseline, training with enriched reports increases AUC from 64.52% to 85.20% and F1 from 36.68% to 66.75%, highlighting the value of breast-specific language supervision.

Building on this enriched supervision, BCHA further improves AUC by 2.81% and F1 by 7.27%, supporting the effectiveness of aligning breast representations with anatomically corresponding de-

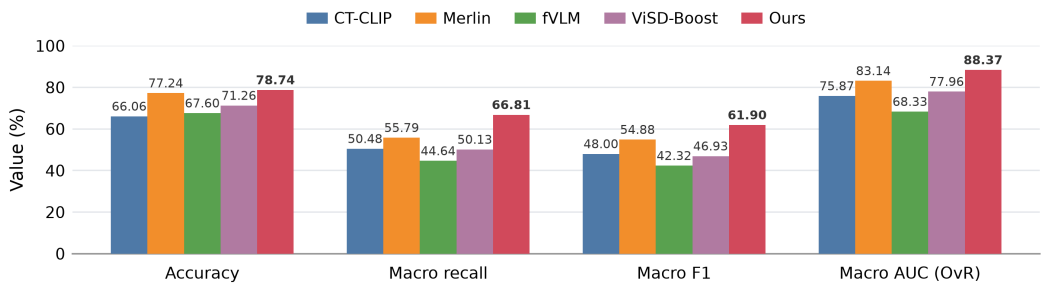


Figure 4: Three-group BI-RADS risk classification on the BreastCT-12K test set (%). Bars compare accuracy, macro recall, macro F1, and macro one-vs-rest AUC; colors denote methods.

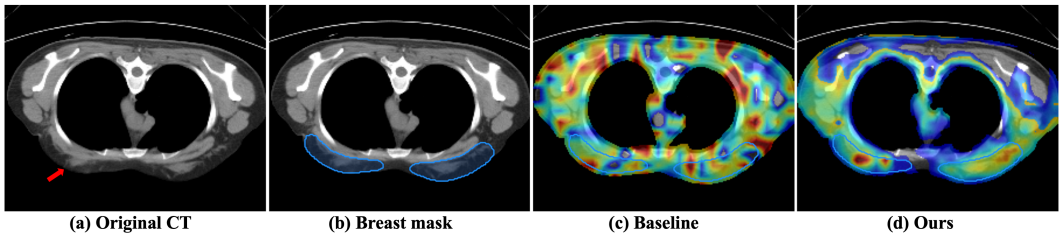
Table 4: Ablation of breast-report enrichment, BCHA, and SEPA for zero-shot diagnosis. Checkmarks indicate enabled components. Best/second-best results are bold/underlined, including ties.

Model	MRI report	BCHA	SEPA	AUC	ACC	Spec	Sens	F1	Prec
Baseline	×	×	×	64.52	71.18	<u>90.89</u>	27.04	36.68	56.99
Breast CT report	×	×	×	80.89	75.43	78.82	67.86	63.03	58.85
BreastCT-12K	✓	×	×	85.20	78.35	81.89	70.41	66.75	63.45
BreastCT-12K+BCHA	✓	✓	×	<u>88.01</u>	<u>83.86</u>	88.04	<u>74.49</u>	<u>74.02</u>	<u>73.55</u>
BreastCT-12K+SEPA	✓	×	✓	86.37	78.43	80.07	**74.75**	68.14	62.61
Full model	✓	✓	✓	**89.19**	**86.54**	**91.80**	**74.75**	**77.41**	**80.27**

scriptions. SEPA alone raises sensitivity from 70.41% to 74.75%, suggesting that finer evidence–patch associations help identify malignant cases. Combining BCHA and SEPA achieves 89.19% AUC, 86.54% accuracy, and 77.41% F1, outperforming either module alone and supporting their complementary roles in learning breast-aware representations.

4.7 Qualitative Visualization

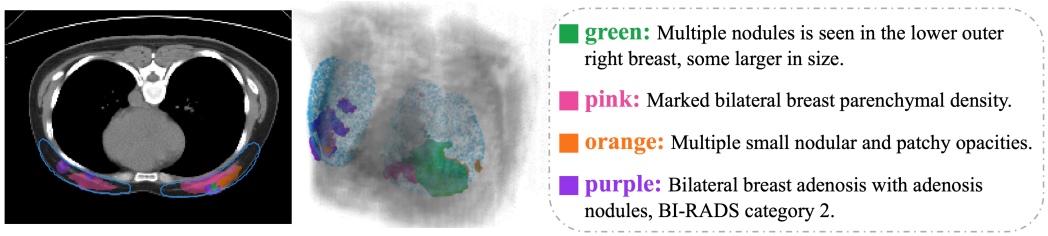
Figure 5 shows a representative text-conditioned similarity map. The general CT baseline responds diffusely across the thorax, while REACH-breast places stronger similarity within the outlined breast regions, including the marked finding. This example visually complements the breast-level alignment gains in Table 4; test-set feature projections are provided in Appendix Figure 9.



(a) Original CT (b) Breast mask (c) Baseline (d) Ours

Figure 5: Breast-text similarity maps for the baseline and REACH-breast. The arrow marks the finding, blue contours outline breasts, and warmer colors indicate higher similarity.

Qualitative Visualization of SEPA. Figure 6 visualizes an example case in slice and 3D views. Blue contours outline breast masks. Nodule evidence selects a unilateral patch cluster, whereas density evidence spans both breasts, illustrating SEPA's anatomically coherent evidence–patch associations.



(a) Text-associated patches (b) 3D patch associations (c) Sentence-level textual evidence

Figure 6: SEPA evidence–patch associations: (a) axial CT view, (b) 3D visualization, (c) sentence-level report evidence. Matching colors link textual evidence to corresponding breast-region patches.

5 Conclusion

In this paper, we presented REACH-breast, a vision–language model for opportunistic breast cancer assessment on routine chest CT. We constructed BreastCT-12K and enriched CT reports with CT-visible breast descriptions curated from temporally matched MRI reports. BCHA connects side-specific and bilateral breast semantics with global representations, while SEPA associates sentence-level evidence with local breast patches without lesion-level annotations. Experiments demonstrate gains in zero-shot diagnosis, bidirectional retrieval, report generation, and BI-RADS classification, with external-center evaluation supporting diagnostic generalization. Additional analyses support complementary module contributions and retention of useful thoracic representations.

AI USE STATEMENT

Generative AI tools assisted with translation, language editing, and feedback on manuscript organization. As described in the data-construction procedure, an LLM also screened and structured MRI-derived breast descriptions, which were checked by human reviewers against the source reports before use. This report-curation step did not create synthetic patient images or replace physician-assigned diagnostic labels. The authors reviewed the AI-assisted content, verified references and experimental results, and take responsibility for the methods, analyses, and final manuscript.

ETHICS STATEMENT

The construction of BreastCT-12K and the use of data from the external centers were approved by the ethics committees of the participating institutions. All data were de-identified before use and handled in accordance with institutional privacy and data protection requirements.

REPRODUCIBILITY STATEMENT

We will publicly release selected implementation code and associated model weights upon acceptance. BreastCT-12K cannot be publicly released due to institutional confidentiality and data-sharing restrictions. The manuscript and appendix materials describe the data construction, training, and evaluation procedures to support reproducibility.

REFERENCES

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, et al. Merlin: A computed tomography vision–language foundation model and dataset. *Nature*, 652:1318–1328, 2026. doi: 10.1038/s41586-026-10181-8.

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

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: 10.3322/caac.21834.

Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, et al. Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature Medicine*, 29: 3033–3043, 2023. doi: 10.1038/s41591-023-02640-w.

Weiwei Cao, Jianpeng Zhang, Yingda Xia, Tony C. W. Mok, Zi Li, Xianghua Ye, Le Lu, Jian Zheng, Yuxing Tang, and Ling Zhang. Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11238–11247, 2024. doi: 10.1109/CVPR52733.2024.01068.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, and Ling Zhang. Boosting vision semantic density with anatomy normality modeling for medical vision–language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025. doi: 10.1109/ICCV51701.2025.02139.

Xin Chen, Mingyan Qiu, Jianpeng Zhang, Yingda Xia, Kai Cao, Weiwei Cao, Lisha Yao, et al. Colorectal cancer detection using noncontrast CT and deep learning: A multicenter and international cohort study. *Annals of Oncology*, 37(8):1144–1156, 2026. doi: 10.1016/j.annonc.2026.04.009.

Yuexi Du, Lihui Chen, and Nicha C. Dvornek. Geometry-guided local alignment for multi-view visual language pre-training in mammography. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15965 of *Lecture Notes in Computer Science*, pp. 299–310. Springer, 2025a. doi: 10.1007/978-3-032-04978-0_29. URL https://doi.org/10.1007/978-3-032-04978-0_29.

Yuexi Du, John A. Onofrey, and Nicha C. Dvornek. Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. In *Information Processing in Medical Imaging*, volume 15830 of *Lecture Notes in Computer Science*, pp. 247–262. Springer, 2025b. doi: 10.1007/978-3-031-96625-5_17. URL https://doi.org/10.1007/978-3-031-96625-5_17.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. doi: 10.48550/arXiv.2403.17834.

Can Hu, Yingda Xia, Zhilin Zheng, Mengxuan Cao, Guoliang Zheng, Shangqi Chen, Jiancheng Sun, et al. AI-based large-scale screening of gastric cancer from noncontrast CT imaging. *Nature Medicine*, 31(9):3011–3019, 2025. doi: 10.1038/s41591-025-03785-6.

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. doi: 10.1109/ICCV48922.2021.00391.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. Visual grounding of whole radiology reports for 3d ct images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224 of *Lecture Notes in Computer Science*, pp. 611–621. Springer, 2023. doi: 10.1007/978-3-031-43904-9_59.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision–language pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15011 of *Lecture Notes in Computer Science*, pp. 80–90. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72104-5_8.

Jingyang Lin, Yingda Xia, Jianpeng Zhang, Ke Yan, Le Lu, Jiebo Luo, and Ling Zhang. CT-GLIP: 3d grounded language–image pretraining with ct scans and radiology reports for full-body scenarios. *arXiv preprint arXiv:2404.15272*, 2024. doi: 10.48550/arXiv.2404.15272.

Philip Müller, Georgios Kaissis, Chen Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, pp. 685–701, 2022. doi: 10.1007/978-3-031-20059-5_40.

Jiyeon Park, Chohee Kim, Yoon Ki Cha, and Myung Jin Chung. Missed and detected incidental breast cancers on contrast enhanced chest CT: Detection rates and CT features. *Diagnostics*, 13 (9):1522, 2023. doi: 10.3390/diagnostics13091522. URL https://doi.org/10.3390/diagnostics13091522.

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, pp. 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

Bowen Shi, Weiwei Cao, Ruifeng Yuan, Wanxing Chang, Wenrui Dai, Hongkai Xiong, Ling Zhang, and Jianpeng Zhang. Disease-centric vision–language pretraining with hybrid visual encoding for 3d computed tomography. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. Also available as arXiv:2606.25546.

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision–language pre-training for enhanced ct image understanding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=K3ZFd7v9rA.

Sarah P. Thomas, Benjamin Wildman-Tobriner, Lasya Daggumati, Lawrence Ngo, Jacob Johnson, Kevin R. Kalisz, Hongyi Zhang, and Tyler J. Fraum. Identifying radiologically significant incidental breast lesions on chest CT: The added value of artificial intelligence. *Current Problems in Diagnostic Radiology*, 2025. doi: 10.1067/j.cpradiol.2025.06.001. URL https://doi.org/10.1067/j.cpradiol.2025.06.001. Online ahead of print.

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. 33536–33548, 2022a.

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. Association for Computational Linguistics, 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. *Nature Communications*, 14: 4541, 2023. doi: 10.1038/s41467-023-40260-7.

Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, and Kayhan Batmanghelich. Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13435 of *Lecture Notes in Computer Science*, pp. 658–668. Springer, 2022. doi: 10.1007/978-3-031-16443-9_63.

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 Machine Learning Research*, volume 182, pp. 2–25. PMLR, 2022.

A Additional Experimental Details

A.1 Evaluation Metrics

We compare the performance of different methods on zero-shot diagnosis, radiology report generation, BI-RADS classification, and image-text retrieval. Zero-shot diagnosis is patient-level binary classification of malignant versus non-malignant cases. We report the area under the ROC curve (AUC), accuracy (ACC), specificity (Spec), sensitivity (Sens), precision (Prec), and F1-score. For the binary diagnosis results in Table 1, decision thresholds are selected directly on the test sets following CT-CLIP's ROC-based procedure (Hamamci et al., 2024). For each model, a single threshold is applied consistently across the three external centers, without center-specific threshold optimization. Accordingly, accuracy, sensitivity, specificity, precision, and F1 are reported at test-set-selected operating points, rather than at independently calibrated thresholds. AUC is computed from prediction scores and is threshold-independent.

For the radiology report generation task, we employ both lexical-overlap metrics and natural language generation metrics for model evaluation, such as ACC, precision, recall, and F1 computed from token overlap between generated and reference reports, together with BLEU-4, METEOR, and ROUGE-L for text quality. The reference text combines the CT *Findings* and *Impression* with the curated MRI-derived breast sections. The evaluation script lowercases each reference and generated report, replaces punctuation with spaces, and splits the text on whitespace. ACC, precision, recall, and F1 measure lexical overlap; no abnormality labeler or predefined disease-label set is used for these four scores. Let n_r and n_g be the reference and generated token counts, and let $o = \sum_w \min(c_r(w), c_g(w))$ count matched token occurrences, where $c_r(w)$ and $c_g(w)$ are the counts of token w. Per-report scores are

$$P = \frac{o}{n_g}, \qquad R = \frac{o}{n_r}, \qquad F1 = \frac{2PR}{P+R}, \qquad \mathrm{ACC}_{\mathrm{tok}} = \frac{o}{\max(n_r, n_g)}. \tag{9}$$

Scores are set to zero when either report is empty; F1 is zero when $P + R = 0$. We compute each score separately for every report pair and take its arithmetic mean across reports, rather than pooling all token counts. Consequently, the mean F1 need not equal the harmonic mean of the reported mean precision and recall. Table 2 also reports sentence-level cumulative BLEU-4 with a brevity penalty and no smoothing, ROUGE-L based on the longest common subsequence ($\beta = 1.2$), and an exact-token METEOR-style score with a fragmentation penalty. The local METEOR implementation does not use stemming or synonym matching. These scores are likewise averaged over reports and expressed as percentages; they assess textual agreement rather than clinical-label accuracy.

For BI-RADS-based evaluation, BI-RADS category labels were assigned by physicians during dataset collection. We aggregate the original assessment categories into three study-defined groups: low-risk (BI-RADS 1–3), intermediate-risk (4A), and high-risk (4B or higher). (BreastCT-12K contains no BI-RADS 0 cases.) We report accuracy, macro precision, macro recall, macro F1, weighted F1, and one-vs-rest macro and weighted AUC.

For the bidirectional image-text retrieval task, we use Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP). R@k in CMC measures the probability of correctly identifying the target instance within the top-k retrieval results, and mAP quantifies overall retrieval precision by averaging precision across recall thresholds.

For general thoracic recognition, we follow the per-finding evaluation and macro-aggregation protocol of fVLM (Shui et al., 2025): balanced accuracy is the mean of sensitivity and specificity for each finding, and each metric is then averaged equally over the 16 selected findings (Appendix D).

A.2 Implementation Details

CT volumes are first resampled to a voxel spacing of $1.5 \times 1.5 \times 3$ mm, then padded as needed and center-cropped to $224 \times 224 \times 160$ voxels. Hounsfield Unit values are clipped to $[-1000, 1000]$ and linearly normalized to $[-1, 1]$. To obtain breast-region masks, we train an nnU-Net Isensee et al. (2021) on 500 CT cases using breast-region masks manually annotated by physicians as ground truth, and apply the trained model to the remaining cases. The masks are resampled to the same voxel spacing using nearest-neighbor interpolation and undergo the same padding and center cropping as their corresponding CT volumes to maintain spatial alignment. The processed masks are then

converted into patch-level coverage scores. A patch is considered valid for breast-region alignment if its mask coverage is at least 0.1.

Patient-level data separation. The approximately 200,000 CT–report pairs used for general pretraining are patient-disjoint from BreastCT-12K and the external breast evaluation cohort. The 500 CT cases with physician-annotated breast masks were collected from a separate hospital and share no patients with the general pretraining corpus, BreastCT-12K, the external evaluation cohorts, or the public datasets used in this study. Thus, neither general pretraining nor breast-segmentation training includes patients from the breast evaluation cohorts.

Our visual encoder follows a 3D ViT-base architecture with 12 transformer blocks and hidden dimension 768. The patch size is $8 \times 8 \times 8$ in depth, height, and width. We initialize the model from our generic 3D CT-CLIP checkpoint trained on approximately 200K chest CT-report pairs. For text encoding, we use the BiomedVLP-CXR-BERT model. The maximum text length is set to 512 tokens. For breast-enhanced training, CT reports and CT-visible MRI evidence are formatted as structured text, where findings and impressions are split into sentence-level segments using the `[SEP]` token.

During breast-enhanced training stage, we train the last two image transformer blocks, the last two text encoder blocks, the projection layers, and the additional breast-region pooling modules. The model is optimized with AdamW using a learning rate of 2.5×10^{-5} and weight decay of 0.05 on 4 NVIDIA A100 GPUs. We train for 20 epochs with a batch size of 32 per GPU using mixed precision. The loss weights are set to 1.0 for global alignment, 1.0 for left/right breast alignment, 1.0 for bilateral breast alignment, 1.0 for global-bilateral cross-level alignment, and 0.5 for SEPA. To avoid unstable local grounding at the beginning of training, SEPA is activated after three warm-up epochs. We use the checkpoint from the final (20th) epoch for evaluation.

For the adapted-model comparisons, all methods use the same training split and are evaluated on the same held-out test split. We additionally evaluate off-the-shelf vision–language models without breast-specific adaptation on the internal test set. Zero-shot diagnosis is performed by comparing CT image embeddings with class-specific textual prompts, as shown in A.3. Retrieval is evaluated by cosine similarity between CT and report embeddings. For report generation, each model is equipped with a model-specific text decoder, which is fine-tuned on the training split while keeping the comparison protocol identical across models.

Report-generation training. We pair each model's image encoder with the same BLIP-style `XBertLMHeadDecoder`, a BERT-base autoregressive decoder with cross-attention to visual tokens. A separate decoder is fine-tuned on the BreastCT-12K training split for each model checkpoint, allowing it to learn to decode that model's visual representations into the enhanced reference reports. Following the frozen and fine-tuned protocols of fVLM (Shui et al., 2025), the frozen setting updates the decoder while keeping the image encoder fixed, whereas the fine-tuned setting also updates the image encoder. Both settings are evaluated on held-out test reports. The inference script loads the corresponding fine-tuned decoder weights and generates reports autoregressively using nucleus sampling. The baseline in Appendix G is the general CT vision–language baseline used to initialize REACH-breast, paired with its separately fine-tuned report decoder.

Report curation and manual review. Following LLM-based screening, the breast descriptions underwent five rounds of review and correction:

1. **Removal of MRI-specific descriptions.** Ineligible sentences describing enhancement, including absent enhancement, linear enhancement, and marked background parenchymal enhancement, were removed. Enhancement-curve descriptions were also excluded.
2. **Formatting and terminology correction.** Measurement separators were standardized from $\times$ to an asterisk. Redundant measurement expressions and typographical errors were corrected. Semicolons in the Impression fields were replaced with periods, and line breaks and ductal-dilatation descriptions were removed from the Findings fields.
3. Context reorganization and deduplication. Longitudinal comparison phrases, such as unchanged or increased size, were moved from unilateral fields to the bilateral field to retain comparison context. Repeated Impression statements, contradictory symmetry descriptions, and redundant wording were removed.

4. Correction of processing artifacts. The processed text was reviewed for errors introduced during earlier replacements, including duplicated terms, malformed boundary descriptions, and unintended word substitutions. These artifacts were corrected to restore the intended wording.

5. Laterality verification. Entries containing contralateral findings were reviewed individually. Statements were separated and reassigned to the appropriate left- or right-breast field.

Short vascular descriptions were additionally excluded from the curated breast text. Repeated-character cleanup preserved legitimate medical expressions. These steps refined the report text without changing the physician-assigned diagnostic labels.

A.3 Zero-Shot Class Prompts

For prompt-based malignancy classification, we encode the following two fixed class descriptions and compare each CT's global image embedding with them. The same prompts are used for every patient, without patient-specific MRI reports at inference.

Non-malignant. "Findings: No obvious suspicious abnormality is seen in either breast. The skin surfaces are smooth, no nipple retraction is seen, no abnormal axillary lymph nodes are seen, and the breasts are symmetric. Impression: Normal or low-risk breast finding."
Malignant. "Findings: Irregular breast mass with indistinct or spiculated margins and possible enlarged axillary lymph nodes. Impression: Mass in the breast; consider malignant tumor, possible invasive ductal carcinoma."

A.4 Model-specific Case Availability

The breast-region evaluation pipelines for fVLM and ViSD-Boost require valid breast masks. A small number of examinations lack usable masks, for example after mastectomy, and are skipped by these pipelines. Their evaluable case sets can therefore differ from those of models that do not require masks at inference. The BCHA component-ablation pipeline likewise skips cases when required masks are unavailable, so its metrics are computed over the successfully processed cases rather than necessarily all 1,270 internal test patients. These exclusions reflect input availability rather than a separately sampled test cohort, and should be considered when comparing the reported results. This pipeline-level filtering is distinct from the model's mask-free global-embedding inference.

A.5 Symmetric Contrastive Loss

For normalized image embeddings $V = \{v_i\}$ and text embeddings $T = \{t_i\}$ on valid patient set $\mathcal{I}$, the symmetric CLIP loss uses learned temperature τ and diagonal matched pairs:

$$\ell_\tau(V, T; \mathcal{I}) = -\frac{1}{2|\mathcal{I}|} \sum_{i \in \mathcal{I}} \Bigg[\log \frac{\exp(v_i^\top t_i/\tau)}{\sum_{j \in \mathcal{I}} \exp(v_i^\top t_j/\tau)} + \log \frac{\exp(t_i^\top v_i/\tau)}{\sum_{j \in \mathcal{I}} \exp(t_i^\top v_j/\tau)} \Bigg], \tag{10}$$

A.6 External Cohorts and Evaluation

The external evaluation included 5,526 patients from three centers, with 1,913 malignant and 3,613 non-malignant cases (Table 5). All three cohorts were held out from model training and hyperparameter selection. AUROC, accuracy, F1-score, and precision were computed separately within each center and averaged with equal weight across centers. For each model, the three external centers share a single decision threshold. Metrics are computed separately within each center and then averaged with equal weight across centers.

Table 5: Patient counts across the three external centers.

Center	Malignant	Non-malignant	Total
Center 1	957	330	1,287
Center 2	846	340	1,186
Center 3	110	2,943	3,053
Total	1,913	3,613	5,526

B Off-the-shelf Diagnosis Results

Table 6 reports six diagnostic metrics for four off-the-shelf models without breast-specific adaptation in the BreastCT-12K internal evaluation. CT-CLIP has the highest AUC (58.21%), whereas ViSD-Boost has the highest accuracy (62.42%) but only 17.86% sensitivity. fVLM and Merlin attain high sensitivity (100.00% and 97.70%) with low specificity (0.35% and 4.40%). Threshold-dependent metrics use the threshold maximizing Youden's index on the evaluation data and are therefore descriptive; AUC is threshold-independent. Figure 3 shows the corresponding ROC curves before and after breast adaptation. Appendix A.4 explains the exclusion of cases without usable breast masks in mask-dependent evaluation pipelines.

Table 6: Off-the-shelf zero-shot diagnosis in the BreastCT-12K internal evaluation (all values in %). No model receives breast-specific adaptation. AUC is threshold-independent; the corresponding ROC curves are shown in Figure 3. Accuracy, specificity, sensitivity, F1, and precision use the Youden-index threshold selected on the evaluation data.

Model	AUC	ACC	Spec	Sens	F1	Prec
CT-CLIP	58.21	62.10	69.91	44.90	42.51	40.37
fVLM	47.21	31.45	0.35	100.00	47.66	31.28
Merlin	46.04	33.52	4.40	97.70	47.85	31.68
ViSD-Boost	48.12	62.42	82.64	17.86	22.88	31.82

C Detailed BI-RADS Classification Results

Table 7 gives the complete results for the three physician-assigned BI-RADS groups defined in Appendix A.1. REACH-breast leads all seven metrics, including 78.74% accuracy, 66.81% macro recall, 61.90% macro F1, and 88.37% macro one-vs-rest AUROC. Its macro recall exceeds Merlin's by 11.02 points (66.81% versus 55.79%). Macro precision, recall, F1, and AUROC are computed for each risk group and then averaged equally across the three groups, rather than interpreted as fractions of correctly classified patients.

Table 7: Three-group BI-RADS classification on the BreastCT-12K internal test set (%). The groups are BI-RADS 1–3, 4A, and 4B or higher. AUROC is computed one-versus-rest; macro averages weight groups equally and weighted averages use class support. Best and second-best scores per metric are bold and underlined.

Method	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1	Macro AUROC	Weighted AUROC
CT-CLIP	66.06	48.19	50.48	48.00	69.23	75.87	79.38
Merlin	<u>77.24</u>	54.24	<u>55.79</u>	<u>54.88</u>	<u>77.74</u>	<u>83.14</u>	<u>87.01</u>
fVLM	67.60	40.92	44.64	42.32	66.99	68.33	70.19
ViSD-Boost	71.26	<u>61.12</u>	50.13	46.93	71.04	77.96	80.93
Ours	**78.74**	**64.05**	**66.81**	**61.90**	**81.78**	**88.37**	**92.11**

D PRESERVATION OF GENERAL THORACIC DISEASE RECOGNITION

We assess whether breast-enhanced training preserves general thoracic recognition on the held-out CT-RATE test split. Following fVLM's per-finding evaluation and macro-aggregation protocol (Shui et al., 2025), we evaluate the 16 findings listed in Figure 7 and average their scores equally. Balanced accuracy is computed as the mean of sensitivity and specificity within each finding; weighted F1 uses the supports of its positive and negative classes, and precision refers to the positive class. Threshold selection follows CT-CLIP's ROC-based procedure (Hamamci et al., 2024), which selects an operating point closest to the upper-left corner of the ROC plot. AUC is threshold-independent.

Table 8 combines results from two sources. For CT-CLIP, CT-VocabFine, and CT-LiPro, we select these same 16 findings from the per-disease results in the CT-CLIP appendix and recompute their arithmetic means (Hamamci et al., 2024). BIUD and Merlin values are transcribed from their respective published reports (Cao et al., 2024; Blankemeier et al., 2026). These methods are not rerun here. REACH-breast attains 75.13% AUC, 74.57% balanced accuracy, 75.26% weighted F1, and 38.30% precision, the highest values among the tabulated references. Its AUC exceeds the recomputed CT-CLIP mean by 3.13 percentage points. The published references provide context rather than a controlled comparison under identical training and inference conditions.

This analysis focuses on retaining whole-volume representations after breast adaptation. In contrast, fVLM uses anatomy-conditioned image–text matching, and ViSD-Boost additionally uses anatomy-specific normality modeling to enhance abnormal visual features (Shui et al., 2025; Cao et al., 2025). Their published thoracic results therefore reflect dedicated regional inference pipelines, whose absolute performance does not isolate retention in our global-embedding pathway. We exclude them from this limited reference comparison and assess both methods in the breast experiments. For a direct retention assessment, Figure 8 compares REACH-breast with its own general CT baseline before breast-enhanced training on each finding.

Table 8: General thoracic recognition on CT-RATE (%). REACH-breast is evaluated on the 16 findings in Figure 7; scores are averaged equally across findings. Balanced accuracy averages sensitivity and specificity, and weighted F1 uses the two class supports within each finding. CT-CLIP and its variants are recomputed from the same 16 per-disease entries in the published appendix; BIUD and Merlin are transcribed from their source papers. Best and second-best tabulated values per metric are bold and underlined.

Metric	CT-VocabFine	CT-LiPro	CT-CLIP	BIUD	Merlin	Ours
AUC	<u>74.81</u>	<u>74.81</u>	72.00	71.30	72.80	**75.13**
Balanced Accuracy	<u>70.19</u>	68.12	66.06	68.10	67.20	**74.57**
Weighted F1	<u>73.38</u>	71.75	69.94	71.60	70.90	**75.26**
Prec	<u>35.19</u>	33.81	32.00	33.80	33.70	**38.30**

D.1 Per-Disease Results

Figure 7 expands the REACH-breast aggregate results in Table 8 to all 16 findings. Pleural effusion has the highest AUC (96.09%), balanced accuracy (90.42%), and weighted F1 (91.21%), whereas lung nodule has the lowest values (59.31%, 56.47%, and 55.92%, respectively). Positive-class precision ranges from 20.21% for interlobular septal thickening to 59.85% for emphysema. These differences show why the aggregate mean should be read alongside the per-disease results.

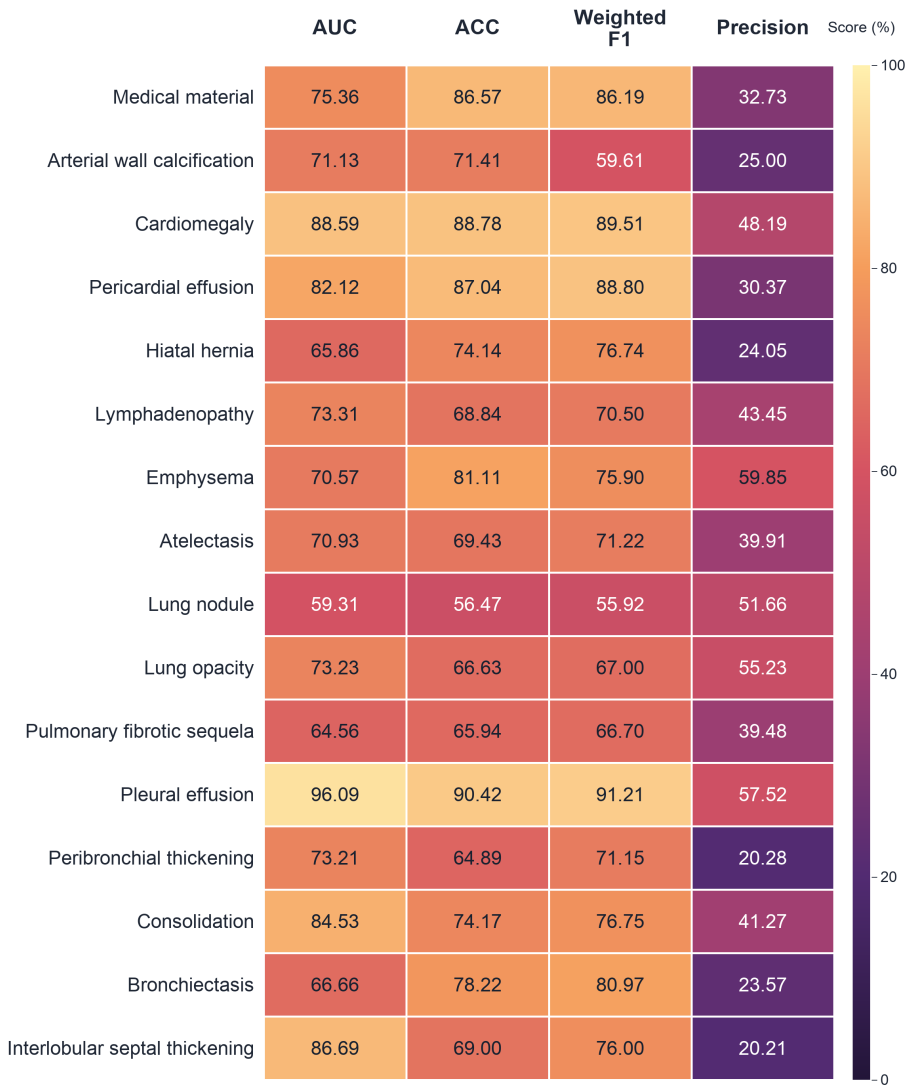


Figure 7: Per-disease results of REACH-breast on the held-out CT-RATE test split (%). Rows are the 16 thoracic findings; columns show AUC, balanced accuracy, class-support-weighted F1, and positive-class precision. Threshold-dependent metrics follow the CT-CLIP threshold-selection procedure; AUC is threshold-independent. Each cell gives the numeric score, and color indicates its magnitude on a shared 0–100% scale.

D.2 PER-DISEASE ROC COMPARISON

Figure 8 shows that retention varies across findings. REACH-breast remains close to the baseline for consolidation (84.5% versus 84.8% AUC), pleural effusion (96.1% versus 97.1%), and cardiomegaly (88.6% versus 90.4%). It improves lung opacity (73.2% versus 65.0%) and interlobular septal thickening (86.7% versus 86.2%). Conversely, arterial wall calcification shows a larger decline (71.1% versus 88.4%). These disease-specific patterns support preservation of useful general thoracic representations, while showing that breast adaptation does not retain every finding equally well. Percentages in this comparison follow the rounded AUCs displayed in the plot; the more precise REACH-breast scores are given in Figure 7.

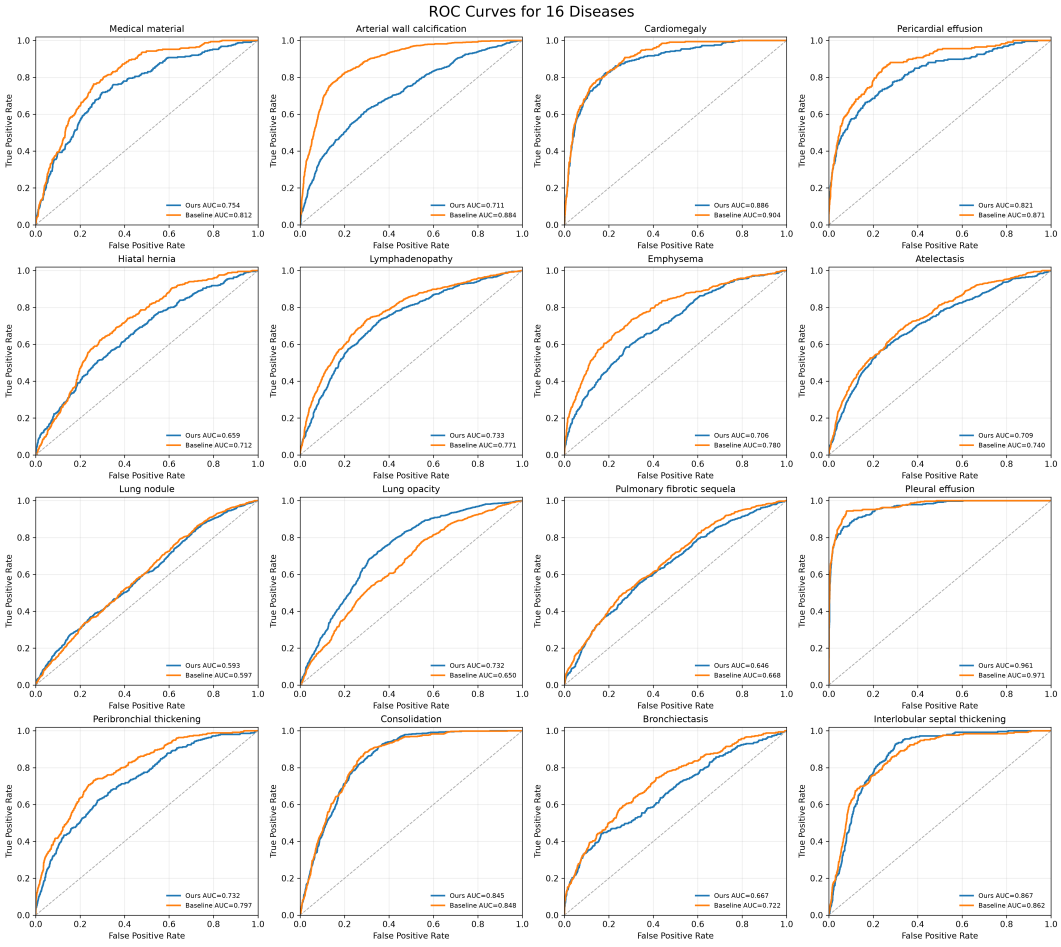


Figure 8: Per-disease ROC curves for the 16 CT-RATE thoracic findings. Each panel compares the general CT baseline before breast-enhanced training (orange) with REACH-breast (blue, labeled Ours); the legend gives the corresponding AUCs on the 0–1 scale, rounded to three decimal places. The horizontal and vertical axes show false positive and true positive rates, respectively. Dashed diagonals denote chance-level discrimination.

E TEST-SET IMAGE FEATURE VISUALIZATION

Figure 9 compares two-dimensional t-SNE projections of test-set image features from the random, baseline, and REACH-breast settings. The random representation shows substantial overlap between non-malignant and malignant cases. The baseline representation shows some local grouping, whereas the REACH-breast representation shows a more distinct concentration of malignant cases. This visualization is qualitative: t-SNE can distort distances and cluster structure, so diagnostic performance is assessed using the quantitative results in Section 4.

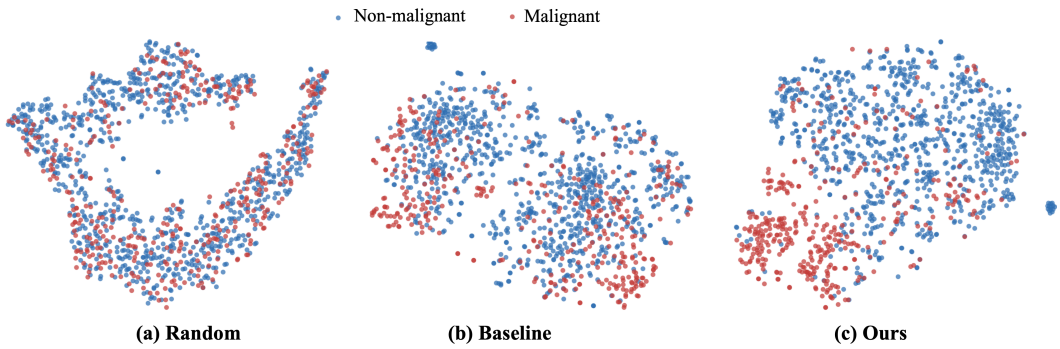


Figure 9: t-SNE projections of test-set CT image features from (a) random initialization, (b) the general CT baseline, and (c) REACH-breast after breast-enhanced training. Each point represents one case; blue indicates non-malignant and red indicates malignant. The projection is qualitative and is not used to compute classification metrics.

F Ablation of BCHA Loss Components

We examine the contributions of the three BCHA objectives by starting from the *BreastCT-12K* configuration in Table 4: training with breast-enriched reports and global CT–text alignment, without BCHA or SEPA. We then add the side, bilateral, and cross losses individually and in combination, keeping SEPA disabled throughout. The side objective includes both left- and right-breast branches; the bilateral objective aligns the combined breast region with its bilateral description; and the cross objective connects global images with bilateral texts and bilateral regions with global texts. Enabled BCHA objectives are summed without additional internal weighting coefficients. Cases lacking masks required by the ablation pipeline are skipped during processing; results are computed over the successfully processed cases (Appendix A.4).

Table 9: Ablation of BCHA loss components for internal malignancy classification (%). All configurations use the breast-enriched supervision of the BreastCT-12K row in Table 4 and retain global alignment; SEPA is disabled. A check mark indicates an enabled loss and a cross indicates a disabled loss. Side comprises both unilateral branches. The first row reproduces the starting configuration from Table 4; subsequent rows add individual objectives or their combinations. Best and second-best scores per metric are bold and underlined.

Side	Bilateral	Cross	AUC	ACC	Spec	Sens	F1	Prec
×	×	×	85.20	78.35	81.89	70.41	66.75	63.45
✓	×	×	86.33	78.90	81.44	73.21	68.17	63.78
×	✓	×	86.13	79.13	81.89	72.96	68.34	64.27
×	×	✓	87.42	82.36	86.45	73.21	71.93	70.69
✓	✓	×	87.34	80.95	83.83	<u>74.49</u>	70.70	67.28
✓	×	✓	87.01	<u>83.07</u>	<u>86.67</u>	**75.00**	<u>73.23</u>	<u>71.53</u>
×	✓	✓	<u>87.68</u>	82.68	86.33	<u>74.49</u>	72.64	70.87
✓	✓	✓	**88.01**	**83.86**	**88.04**	<u>74.49</u>	**74.02**	**73.55**

Each individual objective improves AUC and F1 over the starting configuration (Table 9). Side and bilateral alignment reach 86.33% and 86.13% AUC, respectively, supporting the value of matching anatomical breast views to their descriptions. Cross alignment is the strongest single objective, reaching 87.42% AUC and 71.93% F1, consistent with the benefit of connecting breast-specific semantics to global representations.

The combinations show complementary contributions. Side plus bilateral alignment raises AUC to 87.34%, while adding cross alignment to the bilateral objective reaches 87.68%. Side plus cross gives the highest accuracy and F1 among two-objective variants, although it does not give their highest AUC. Combining all three objectives yields the best values on five metrics: 88.01% AUC, 83.86% accuracy, and 74.02% F1, with 73.55% precision and 88.04% specificity. Relative to the starting configuration, the complete BCHA objective improves AUC by 2.81 points and F1 by 7.27 points. These results support jointly preserving side-specific information, modeling bilateral context, and linking breast regions with whole-volume semantics.

G Qualitative Report Generation Examples

Figures 10–12 compare three cleaned reference reports with outputs from the baseline and REACH-breast. The baseline reports focus on thoracic structures and omit breast descriptions in all three examples. REACH-breast generates breast-specific findings and impressions, including breast symmetry, nodules, and a unilateral mass. This complements the aggregate evaluation in Table 2 by illustrating its ability to express breast-related information in report form. All report statements are retained as supplied; differences from the reference and internal inconsistencies are discussed below each example.

G.1 Example 1: Breast Symmetry and Negative Findings

Reference report (GT)

Findings: The thorax is symmetric. No significant osseous or soft-tissue abnormality is identified. The lungs are clear bilaterally. The bronchovascular markings are normal in caliber and distribution. The tracheobronchial tree is patent. The hilar structures are unremarkable. The mediastinum is midline. No significant abnormality is identified in the structure, position, or size of the heart or great vessels. No significantly enlarged mediastinal lymph nodes are identified. No significant pleural or pleural-space abnormality is present bilaterally. The breasts are grossly symmetric, with prominent bilateral fibroglandular tissue. No focal soft-tissue breast lesion, skin or areolar thickening, nipple retraction, or enlarged axillary lymph nodes is identified.

Impression: No significant abnormality on HRCT of the lungs. No suspicious CT-visible breast abnormality. BI-RADS category 2.

Baseline

Findings: The thorax is symmetric. The mediastinum is midline. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. A small nodule is present in the right upper lobe (IM: 12), with a long-axis diameter of approximately 8 mm. A tiny nodule is present in the left lower lobe (IM: 38). No significant pleural thickening or chest wall abnormality is identified.

Impression: Small nodule in the right upper lobe and tiny nodule in the left lower lobe.

REACH-breast (Ours)

Findings: The thorax is symmetric. The mediastinum is centered, and no enlarged mediastinal lymph nodes are seen. The trachea and main bronchi are patent. Hilar structures are unremarkable. Cardiac size and contour are within normal limits. Bronchovascular markings are normal bilaterally, without focal pulmonary nodules or linear opacities. No pleural abnormality is identified. The breasts are symmetric with prominent fibroglandular tissue. No focal breast soft-tissue nodule is seen. There is no skin or areolar thickening, nipple retraction, or axillary lymphadenopathy.

Impression: Multiple bilateral breast nodules. No significant pulmonary abnormality on HRCT. No suspicious CT-detectable breast lesion. BI-RADS category 2.

Figure 10: Report-generation example 1. The reference describes symmetric breasts without a suspicious focal abnormality. REACH-breast reproduces several breast findings omitted by the baseline. The complete cleaned reports are shown, including the contradictory mention of bilateral breast nodules in the REACH-breast impression.

REACH-breast captures breast symmetry, prominent fibroglandular tissue, and the absence of skin changes, nipple retraction, or axillary lymphadenopathy, consistent with the reference. The baseline omits these findings and instead describes pulmonary nodules absent from the reference. Although REACH-breast reproduces BI-RADS 2, its impression introduces multiple bilateral breast nodules despite denying a focal nodule in the findings. This example illustrates improved breast-content coverage together with a remaining consistency error.

G.2 Example 2: Right-Breast Nodules and Postoperative Findings

Reference report (GT)

Findings: The thorax is symmetric. The mediastinum is midline. The trachea and main bronchi are patent. A calcified left hilar lymph node is present. The cardiac silhouette is unremarkable. A punctate calcification and a fibrotic focus are present in the left lower lobe. No significant pleural thickening or chest wall abnormality is identified. Low-attenuation and gas foci are present in the right breast. Postoperative changes are present in the upper outer right breast, including focal skin irregularity and surrounding fat stranding. Focal architectural distortion is also present in this region. Two well-circumscribed nodules are present in the upper medial and lateral right breast, measuring approximately 1.3 × 0.6 cm and 0.6 × 0.7 cm, respectively. Enlarged right axillary lymph nodes are present, with no enlarged left axillary lymph nodes.

Impression: Calcification and fibrotic focus in the left lower lobe. Postoperative changes following right breast mass excision. Two upper right breast nodules, BI-RADS 3.

Baseline

Findings: The thorax is symmetric. The mediastinum is midline. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. The lungs are clear bilaterally. No significant pleural thickening or chest wall abnormality is identified.

Impression: No significant abnormality on noncontrast chest CT.

REACH-breast (Ours)

Findings: The thorax is symmetric. The mediastinum is midline. No significantly enlarged mediastinal lymph nodes are identified. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. A small subpleural nodule measuring less than 0.6 cm is present in the left lower lobe. No significant pleural thickening is identified. A nodular high-attenuation opacity is present in the upper medial of the right breast. Scattered small soft-tissue nodules are present in both breasts. No breast skin thickening, or enlarged axillary or internal mammary lymph nodes are identified. No chest wall abnormality is present.

Impression: Inflammatory opacity in the left lower lobe. Calcified nodule in the lower of the left breast and calcification in the upper imedial of the right breast. Multiple small bilateral breast nodules.

Figure 11: Report-generation example 2. The reference includes postoperative right-breast changes, two right-breast nodules, and right axillary lymphadenopathy. The baseline omits breast findings, whereas REACH-breast generates breast nodule descriptions, including a right-sided location. Its output does not fully recover the reference findings. Wording is preserved from the supplied cleaned reports.

REACH-breast produces a nodular description in the upper medial right breast and includes breast findings in the impression, whereas the baseline reports no significant chest abnormality. This demonstrates breast-specific content generation, but the match is incomplete: the output omits the postoperative changes and reference nodule measurements, denies the enlarged right axillary nodes, and introduces bilateral nodules and calcifications not described in the reference. The example therefore supports improved attention to breast content without establishing complete recovery of lesion attributes.

G.3 EXAMPLE 3: LEFT-BREAST MASS

Reference report (GT)

Findings: The thorax is symmetric. The mediastinum is midline. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. The lungs are clear bilaterally. No significant pleural thickening or chest wall abnormality is identified. Scattered punctate high-attenuation foci are present in the left breast. Focal skin thickening and an irregular soft-tissue lesion measuring approximately 6.0 × 1.5 cm are present in the upper inner quadrant of the left breast. No enlarged axillary lymph nodes are identified.

Impression: Scattered small calcifications in the left breast. Segmental lesion in the upper inner quadrant of the left breast, BI-RADS 4B.

Baseline

Findings: The thorax is symmetric. The mediastinum is midline. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. Patchy ill-defined opacities are present in the lingular segment of the left upper lobe and the left lower lung. No significant pleural thickening or chest wall abnormality is identified.

Impression: Inflammatory opacities in the lingular segment of the left upper lobe and the left lower lung.

REACH-breast (Ours)

Findings: The mediastinum is midline. The trachea and main bronchi are patent. The bilateral hila are not enlarged. The heart is normal in size and contour. The bilateral bronchovascular markings are normal. A small calcification is present in the left upper lobe. The left inner lobe is prominent. No significant pleural thickening is identified. A soft-tissue mass is present in the left breast. No breast skin thickening, nipple retraction, or enlarged axillary lymph nodes are identified.

Impression: Left breast mass with mildly irregular enlargement of the upper inner lobe. Multiple bilateral breast nodules, BI-RADS 3.

Figure 12: Report-generation example 3. REACH-breast mentions a left-breast mass in both findings and impression, while the baseline describes pulmonary opacities and omits the breast lesion. The reference includes an upper-inner left-breast lesion with skin thickening and BI-RADS 4B; the generated report retains errors in lesion attributes and assigns BI-RADS 3.

The generated left-breast mass agrees with the reference at the level of lesion presence and laterality, an important breast-specific finding absent from the baseline output. However, REACH-breast omits the reference lesion dimensions, denies skin thickening, adds bilateral nodules, and assigns BI-RADS 3 instead of 4B. Across these examples, the gain is the ability to include breast-specific content within chest CT reports; accurate attribute recovery, risk assessment, and consistency between findings and impression remain areas for improvement.