Mammo2Tomo: Adapting Mammo-CLIP to learn text-aligned DBT volume representations

Aya Kassem[1], Shantanu Ghosh[1], Clare B. Poynton[2], and Kayhan Batmanghelich[1]

[1] Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA
ayak@bu.edu
[2] Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA

Abstract. Multimodal foundational models (FMs) for 3D-Digital breast tomosynthesis (DBT) have the potential to improve data efficiency, model explainability, and serve as a building block for breast cancer risk estimation. Although DBT provides richer volumetric information than 2D mammography, developing DBT-specific FMs encounters critical challenges: the unavailability of large multimodal datasets, 2D-to-3D domain shift, and the inherent limitations of 2D FMs to model inter-slice dependencies. This paper proposes Mammo2Tomo, a novel 2-stage adaptor, extending the mammography FM Mammo-CLIP to 3D DBT volumes. Our Stage 1 utilizes a pseudo-labeling strategy to mitigate the slice-level mammography-to-DBT domain shift and better maintain image-text alignment. In Stage 2, we introduce an aggregator anchored by C-view projections to tackle volumetric modeling and capture essential inter-slice dependencies. Experiments on two public DBT datasets show that Mammo2Tomo improves average volume-level AUROC by ∼10% over unadapted Mammo-CLIP and ∼16% over MedSigLIP, demonstrating effective mammography-to-DBT transfer while largely preserving multimodal image–text alignment.

Keywords: DBT · Foundational Models · Vision-Language Models.

Mammography and DBT are routinely used for the screening, diagnosis, and risk estimation of breast cancer, a leading cause of mortality [18]. Recently, several multimodal mammography FMs [8, 2, 5, 4] enhance the data efficiency and interpretability of current computer-aided diagnosis (CAD) systems via image-text alignment. DBT mitigates tissue superposition in 2D mammography, reducing ambiguity and improving diagnostic detail. However, existing FMs are trained solely on 2D mammograms, failing to process 3D DBT volumes. To address this, we propose Mammo2Tomo – an adapter enabling Mammo-CLIP [8] to process DBT volumes while largely preserving its learned image-text alignment.

Developing text-aligned DBT FMs is currently hindered by four challenges. First, generalist FMs like MedSigLIP [17] lack domain-specific sensitivity and

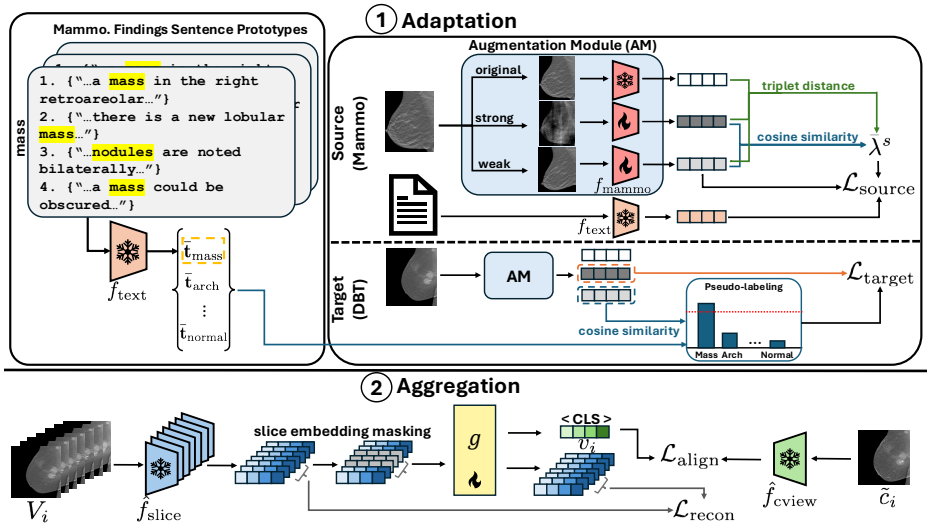

Fig. 1: Overview of Mammo2Tomo. **Stage 1:** Sentence prototypes from source reports provide pseudo-label supervision for adapting $\hat{f}_{\text{mammo}}$ to DBT slices and C-views while f_{text} remains frozen. **Stage 2:** A transformer g aggregates per-slice embeddings from $\hat{f}_{\text{slice}}$ into a volume embedding v_i, supervised by C-view alignment ($\mathcal{L}_{\text{align}}$) and masked slice reconstruction ($\mathcal{L}_{\text{recon}}$).

downsample high-resolution images essential for mammography and DBT. Second, recent DBT-specific models [3] rely on DINO [15], lacking the image-text alignment essential for explainability and downstream report-generation VLMs [7]. Third, training a DBT-specific FM from scratch is constrained by the unavailability of large-scale paired DBT-report corpora; also, it inherently discards the substantial knowledge encoded in 2D mammography FMs. Fourth, applying Mammo-CLIP to individual DBT slices is fundamentally limited by the severe mammography-to-DBT domain shift and ignores crucial inter-slice relationships. Therefore, we formulate an unsupervised domain adaptation (UDA) setup utilizing a source dataset of mammogram-report pairs, with publicly available target DBT datasets [22, 1] lacking text or fine-grained annotations. Despite the extensive UDA literature, most methods focus on task-specific supervised adaptation rather than representation learning. Methods like adversarial domain adaptation [6], Fourier domain adaptation (FDA) [20], or subspace alignment like CORAL [19] are ill-suited for FMs and risk degrading image-text alignment due to catastrophic forgetting.

Contributions. We propose Mammo2Tomo, a two-stage method based on a pseudo-labeling framework for UDA [12]. In the first stage, we adapt the Mammo-CLIP vision encoder to reduce the mammography-to-DBT domain gap, using a dynamic forgetting coefficient to limit catastrophic forgetting and adaptive debiasing to correct pseudo-labeling bias toward frequent classes. In the

second stage, we develop an aggregator that models inter-slice relationships utilizing a masked autoencoder self-supervised learning strategy. To ensure the resulting volumetric representation remains aligned with text, we synthesize an approximate C-view [11] from the DBT slices and use its Mammo-CLIP encoding as a teacher anchor for volumetric representation learning. Our experiments show that both slice- and volume-level representations largely preserve image-text alignment and outperform generic FMs. Furthermore, our approach substantially improves data efficiency and leverages text alignment to directly localize pathologies on DBT slices.

1 Method

Mammo2Tomo consists of two stages: adaptation and aggregation. Stage 1 (adaptation) reduces the domain gap of Mammo-CLIP [8] from 2D mammography (source) to DBT (target) via UDA. This stage exclusively updates the image encoder to process DBT slices while freezing the text encoder. Our adaptation strategy is based on a pseudo-labeling framework and aims to preserve the alignment between image and text after adaptation. Stage 2 (aggregation) aggregates the adapted per-slice representations into a unified, text-aligned volume representation. We formulate the problem to ensure the aggregator can learn the relationship between slices while maintaining text alignment.

Notation. Let $\mathcal{D}_s = \{(x_i^s, t_i^s)\}$ and $\mathcal{D}_t = \{V_i\}$ be the source 2D mammogram–report pairs and the unlabeled target DBT volumes respectively. Each V_i has N_i ordered slices at z-positions $\{p_j\}_{j=1}^{N_i}$ and a synthesized C-view $\tilde{c}_i$. We denote f_{mammo} as Mammo-CLIP's image encoder and f_{text} as its text encoder. $w(\cdot)$ and $s(\cdot)$ denote weak and strong augmentations, respectively. Each breast imaging finding (*e.g.,* mass, calcification) k has a sentence prototype $\bar{t}_k$. Stage 1 yields adapted encoders $\hat{f}_{\text{slice}}$ and $\hat{f}_{\text{cview}}$ from f_{mammo}. Stage 2 learns an aggregator g that takes per-slice representations $\{z_j\}_{j=1}^{N_i} \subset \mathbb{R}^d$ with their z-positions $\{p_j\}$ and produces a volume representation $v_i \in \mathbb{R}^d$ and per-slice outputs $\{\hat{z}_j\} \subset \mathbb{R}^d$.

1.1 Stage 1: Domain Adaptation

Following PadCLIP [12], we formulate an unsupervised domain adaptation (UDA) task to mitigate the distribution shift between f_{mammo} and target DBT images without paired text. We use paired image–text data from the source domain $\mathcal{D}_s$ and unlabeled images from the target domain $\mathcal{D}_t$. Standard pseudo-labels (*e.g.,* "a photo of a mass") ignore the semantic variance of clinical reports. Instead, we average relevant source-domain text embeddings to construct finding-specific (*e.g.,* mass) prototypes $\bar{t}_k$, alongside a normal prototype denoting finding absence.

Objective and total loss. Unlike classical classification-based UDA, we adapt a pre-trained representation model while largely preserving its pre-trained vision–language alignment. To prevent catastrophic forgetting, we optimize the

Table 1: Volume-level AUROC under ZS/LP/FT with mean pooling vs. learned aggregation. Best in **bold**, second underlined.

		OMAMA			BCS-DBT					
		Cancer			Mass			Arch Dist.		
Pool	**Method**	**ZS**	**LP**	**FT**	**ZS**	**LP**	**FT**	**ZS**	**LP**	**FT**
Mean	MedSigLIP	0.51	0.67	0.73	0.57	0.72	0.74	0.48	0.70	0.72
	Mammo-CLIP-UnAdap.	0.49	0.70	0.77	**0.65**	0.72	0.76	0.61	0.71	0.74
	Mammo-CLIP-Adap.	0.57	0.72	0.81	0.58	0.80	0.82	0.65	0.77	0.78
Aggr.	MedSigLIP	0.54	0.63	0.69	0.57	0.70	0.74	0.54	0.71	0.72
	Mammo-CLIP-UnAdap.	0.53	0.72	0.80	0.58	0.76	0.79	0.69	0.74	0.78
	Mammo2Tomo (Ours)	**0.60**	**0.75**	**0.88**	0.60	**0.81**	**0.84**	**0.71**	**0.78**	**0.80**

model end-to-end with a supervised contrastive loss on the source domain and a pseudo-label consistency loss on the target domain. The adaptation loss is:

$$\mathcal{L}_{\text{adapt}} = \bar{\lambda}^s \cdot \mathcal{L}_{\text{source}}(\mathcal{D}_s) + \alpha \cdot \mathcal{L}_{\text{target}}(\mathcal{D}_t). \tag{1}$$

$\bar{\lambda}^s$ is a dynamic knowledge-preservation coefficient: it decreases as catastrophic forgetting increases, effectively lowering the learning rate on the source domain to protect pre-trained weights (detailed below). α is a fixed scaling factor for the target loss. $\mathcal{L}_{\text{source}}$ is the symmetric InfoNCE loss on source pairs $\{(x_i^s, t_i^s)\}_{i=1}^B$:

$$\mathcal{L}_{\text{source}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{e^{u_i \cdot e_i/\sigma}}{\sum_j e^{u_i \cdot e_j/\sigma}} + \log \frac{e^{e_i \cdot u_i/\sigma}}{\sum_j e^{e_i \cdot u_j/\sigma}} \right], \tag{2}$$

where $u_i = f_{\text{mammo}}(x_i^s)$, $e_i = f_{\text{text}}(t_i^s)$, and σ is a learned temperature.

Pseudo-labeling on the target domain. The general idea is to perform pseudo-labeling on the target domain, filtering for confident predictions to act as pseudo-ground-truth supervision. Each target image x_i^t is passed through the augmentation module (AM), producing a weak view $w(x_i^t)$ and a strong view $s(x_i^t)$. The weak-view representations are classified against the sentence prototypes $\{\bar{t}_k\}$ via cosine similarity followed by softmax, producing a soft pseudo-label distribution q_i over findings. Since the weak augmentation stays close to the real image, these predictions are reliable enough to serve as pseudo-labels. The consistency loss is:

$$\mathcal{L}_{\text{target}} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\big[\max(q_i) \geq \tau\big] \, \text{CE}(p(y_i \mid s(x_i^t)), \, q_i), \tag{3}$$

where B is the batch size; $\mathbb{1}[\cdot]$ is an indicator activating the loss only when the pseudo-label confidence satisfies $\max(q_i) \geq \tau$; CE denotes cross-entropy; q_i is the soft pseudo-label distribution generated from the weakly augmented target image; and $p(y_i \mid s(x_i^t))$ is the model prediction computed analogously to q_i using the strongly augmented target image $s(x_i^t)$.

Table 2: Slice-level UDA of Mammo-CLIP→DBT. Best in **bold**, 2^{nd} underlined. MC and `S1` denote Mammo-CLIP and Stage1.

	OMAMA		BCS-DBT			
	Cancer		Mass		Arch Dist.	
Method	ZS	LP	ZS	LP	ZS	LP
MC-UnAdap.	0.47	0.66	0.69	0.82	0.67	0.90
AdaBN	0.46	0.66	0.65	0.84	0.66	0.92
FDA	0.51	0.66	0.72	0.83	**0.76**	0.95
Lin. CORAL	0.50	0.67	0.69	0.82	0.70	0.93
Adversarial	0.50	0.72	**0.75**	**0.88**	0.67	0.94
Ours (S1)	**0.53**	**0.73**	0.72	0.86	0.74	**0.97**

Table 3: T2I slice retrieval reported as Recall@k (Rk), tolerance ±3. Best in **bold**, 2^{nd} underlined. All models are w/ aggr. MC and ours denote Mammo-CLIP and Mammo2Tomo.

	OMAMA		BCS-DBT			
	Cancer		Mass		Arch.	
Method	R1	R3	R1	R3	R1	R3
Random	.11	.29	.10	.28	.10	.28
MedSigLIP	.01	.03	.13	.26	.11	.19
MC-UnAdap.	.27	.38	.30	.45	.11	.24
Ours	**.32**	**.42**	**.36**	**.49**	**.26**	**.40**

Exploration vs. Forgetting. To ensure exploration, we use both weak and strong augmentations. The weak augmentation $w(\cdot)$ applies standard geometric transforms (flips, affine, elastic) to yield representations similar to the original image. To force out-of-distribution exploration, the strong augmentation $s(\cdot)$ additionally applies Fourier Domain Adaptation (FDA) [20], which swaps low-frequency amplitudes with those of random cross-domain samples using a bandwidth parameter β. As the encoder adapts, it risks catastrophic forgetting of its pre-trained alignment. We quantify this drift using the similarity between the original, weak, and strong representations. We extract the ℓ_2-normalized representations for the original, weak, and strong views as $\tilde{e}^i_o = f_0(x_i)$, $\tilde{e}^i_w = f_{\text{mammo}}(w(x_i))$, and $\tilde{e}^i_s = f_{\text{mammo}}(s(x_i))$ respectively, where f_0 is a frozen copy of the base Mammo-CLIP image encoder. The raw triplet distance is:

$$\lambda^s = 1 - \frac{1}{6B}\sum_{i=1}^{B}\big(\|\tilde{e}^i_w - \tilde{e}^i_o\|_2 + \|\tilde{e}^i_s - \tilde{e}^i_o\|_2 + \|\tilde{e}^i_w - \tilde{e}^i_s\|_2\big). \tag{4}$$

To stabilize this dynamic measurement across batches, a momentum term m^s measures the cosine similarity between the weak and strong augmentations: $m^s = \frac{1}{B}\sum_i \langle \tilde{e}^i_w, \tilde{e}^i_s \rangle$. The final dynamic coefficient $\bar{\lambda}^s$ is updated iteratively during adaptation: $\bar{\lambda}^s \leftarrow m^s \cdot \lambda^s + (1 - m^s) \cdot \bar{\lambda}^s_{\text{prev}}$

Adaptive debiasing. To mitigate inter-class bias—where the model reinforces errors by over-predicting majority classes—we track systematic bias via an exponential moving average of target predictions: $\bar{q} \leftarrow m^t \bar{q} + (1 - m^t)\frac{1}{B}\sum_{i=1}^{B} q_i$. We then debias the soft pseudo-labels: $q_i \leftarrow q_i - \bar{\lambda}^t \log \bar{q}$. Here, target momentum m^t and forgetting coefficient $\bar{\lambda}^t$ are computed analogously to m^s and $\bar{\lambda}^s$, smoothly penalizing bias based on uncertainty. Stage 1 independently adapts slices and C-views, yielding $\hat{f}_{\text{slice}}$ and $\hat{f}_{\text{cview}}$.

1.2 Stage 2: Aggregation

Stage 2 trains an aggregator g to convert the per-slice representations $\{z_j\}_{j=1}^{N_i}$ into a unified volume-level representation v_i. This module models inter-slice de-

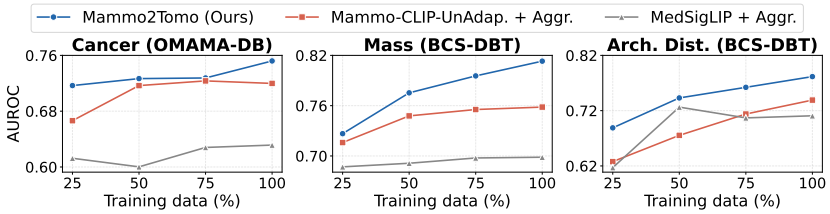

Fig. 2: Data efficiency of MAMMO2TOMO with LP for 25-100% of training data.

pendencies while largely preserving the image–text alignment of $\{z_j\}$. Our aggregator g maps per-slice representations to a volume representation. Representations $\{z_j\}_{j=1}^{N_i}$ from frozen $\hat{f}_{\text{slice}}$ receive positional encodings indexed by their physical z-positions $\{p_j\}_{j=1}^{N_i}$. A learnable `[CLS]` token is prepended:

$$[\hat{z}_{\text{cls}},\ \hat{z}_1,\ \ldots,\ \hat{z}_{N_i}] = g([\texttt{[CLS]},\ z_1,\ \ldots,\ z_{N_i}],\ [\_;\ p_1,\ \ldots,\ p_{N_i}]). \tag{5}$$

The `[CLS]` yields volume representation $v_i = \hat{z}_{\text{cls}}$, with $\{\hat{z}_j\}$ for reconstruction.
Masked reconstruction. Following MAE [9], g reconstructs masked slice representations from the surrounding context. We mask blocks of slices to prevent trivial adjacent-slice interpolation, enforcing long-range 3D modeling. The loss $\mathcal{L}_{\text{recon}} = \text{MSE}(\hat{z}_{\text{masked}}, z_{\text{masked}})$ predicts the original representations at masked positions, where z_{masked} are ground-truth representations from the frozen $\hat{f}_{\text{slice}}$.
C-view alignment. Reconstruction encourages g to understand the 3D structure of DBT slices, but it does not ensure that the aggregated volume representation remains text-aligned. To do so, we anchor v_i using a C-view image – a synthesized 2D mammogram projected from the DBT volume. Since Stage 1 successfully adapts the C-view encoder $\hat{f}_{\text{cview}}$, its resulting representation $c_i = \hat{f}_{\text{cview}}(\tilde{c}_i) \in \mathbb{R}^d$ provides a natural anchor for v_i. Commercial C-view algorithms are proprietary, so we approximate them via high-pass filtered max-intensity projection [11]: V_i is convolved with a $9 \times 9 \times 9$ 3D high-pass spherical filter and max-projected across depth to produce $\tilde{c}_i$. The alignment loss $\mathcal{L}_{\text{align}} = \text{MSE}(v_i, c_i)$ anchors v_i to the text space. The overall aggregation objective combines both signals: $\mathcal{L}_{\text{agg}} = \mathcal{L}_{\text{align}} + \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}}$

2 Experiments

We evaluate whether **(1)** Stage 1 UDA establishes target-domain image–text alignment; **(2)** Stage 2 aggregation yields superior volume representations; **(3)** MAMMO2TOMO remains data-efficient; and **(4)** MAMMO2TOMO preserves pretrained cross-modal semantics.
Datasets. Source domain $\mathcal{D}_s$ comprises 82,624 mammogram–report pairs from the UPMC, EMBED [10], and BUMC datasets, following Mammo-CLIP [8]. We extract finding-specific (*e.g.,* mass) sentence prototypes $\bar{t}_k$ from 27,309 source reports via a rule-based (regex) pipeline with negation detection. Target domain $\mathcal{D}_t$

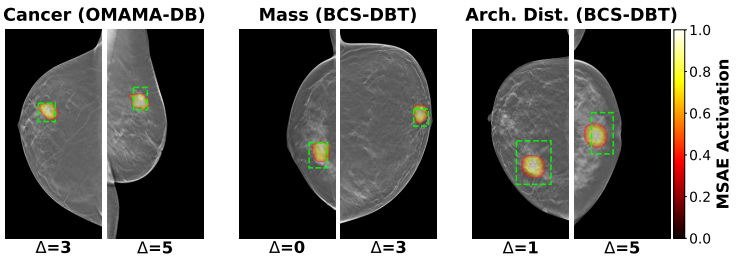

Fig. 3: MSAE-localized T2I slice retrieval. Top-ranked slices per finding are overlaid with activation heatmaps and ground-truth bounding boxes (**green**). Δ indicates the offset from the ground-truth slice.

includes two DBT datasets: BCS-DBT [1] (22,032 volumes from 5,060 patients; 336 mass, 99 architectural distortion boxes) and OMAMA-DB [22] (67,534 volumes from 15,149 patients; 376 cancer cases). OMAMA-DB provides one annotated slice per volume with a cancer/non-cancer bounding box. BCS-DBT offers one bounding box per volume for a subset of positive samples. OMAMA-DB has no standard split, so we build stratified patient-level train/validation/test folds matching RSNA's per-fold prevalence; for BCS-DBT we adopt its provided partition. We report on the held-out test set. Within each split, we keep all positives and subsample negatives to match the prevalence of the RSNA [16] (cancer) and VinDr-Mammo [14] (mass, arch.) benchmarks used by Mammo-CLIP: 2%, 5.5%, and 0.6%.

Experimental Setup. Stage 1 of Mammo2Tomo adopts similar preprocessing of DBT slices and training configurations from Mammo-CLIP [8]. Strong augmentations and weak augmentations utilize FDA ($\beta = 0.03$) and the Mammo-CLIP pre-training transformations, respectively. Following PadCLIP [12], we set $\tau = 0.4$ and $\alpha = 0.5$. We adapt the encoder with AdamW (LR $5{\times}10^{-5}$, weight decay 10^{-4}) and a cosine schedule at a batch size of 4. We precompute the slice and C-view representations from the adapted Stage 1 encoders. Stage 2 of Mammo2Tomo leverages a 2-layer transformer as an aggregator with 8 attention heads, a 512-dimensional hidden state, GELU activations, and a dropout rate of 0.1. We train it for 50 epochs (batch size 16) using the AdamW optimizer with a LR of 10^{-4} and a weight decay of 0.01. We schedule the LR with cosine annealing and a 10-epoch early stopping patience. We mask contiguous 10-slice blocks at a 50% ratio with $\lambda_{\text{recon}} = 0.5$.

Evaluation tasks and baselines. To validate Stage 1 adaptation on image–text alignment, we report zero-shot (ZS) and linear probe (LP) classification for all the labels, comparing against UDA baselines *e.g.,* AdaBN [13], FDA [20], Lin. CORAL [19], and Domain-Adversarial Training [6], applied directly to the Mammo-CLIP image encoder. To evaluate the Stage 2 volume representations, we perform ZS, LP, and fine-tuning (FT) classification for the same three labels, and report AUROC. A limitation is that our results are reported as point estimates, without statistical uncertainty estimates such as confidence inter-

vals. We compare our aggregator against mean pooling across three text-aligned encoders—MedSigLIP (a generalist model), unadapted Mammo-CLIP (Mammo-CLIP-UnAdap.), and our Stage 1-adapted variant (Mammo-CLIP-Adap.); as our goal is text-aligned volume representations, we omit vision-only DBT models (*e.g.,* DBT-DINO [3]) that lack a text encoder for ZS and retrieval. LP is trained for 30 epochs at LR 10^{-3} and FT for 10 epochs at LR $5{\cdot}10^{-6}$. We strictly follow Mammo-CLIP for remaining parameters. Our data efficiency evaluation via LP utilizes 25% to 100% of the target training data to classify the 3 labels. We quantify image–text alignment via text-to-image (T2I) slice retrieval, ranking DBT slices against Stage 1 prototypes $\{\bar{t}_k\}$ by cosine similarity, and report Recall@k, ± 3 tolerance. For spatial localization, we train an MSAE [21] (16,384 neurons, top-128 activation) on adapted features, extracting finding-specific neurons by contrasting activations inside vs. outside ground-truth boxes.

Stage 1 Improves Target-Domain Alignment. Tab. 2 reports slice-level UDA from 2D mammography to DBT. Mammo2Tomo (Stage 1) effectively mitigates the severe 2D-to-3D domain shift, notably outperforming the Mammo-CLIP-UnAdap. in cancer classification with relative AUROC gains of 12.8% (0.53 vs. 0.47) in ZS. Adversarial baselines, though competitive on isolated targets (*e.g.,* BCS-DBT mass), only align global statistics without semantic grounding, whereas Mammo2Tomo provides direct target-domain supervision via finding-specific pseudo-labeling. Furthermore, our adaptive weighting is designed to reduce catastrophic forgetting and better retain transferable knowledge than standard UDA.

Stage 2 Improves Volume Representations. Tab. 1 shows Mammo2Tomo outperforms all baselines on volume-level classification. For ZS cancer, Mammo2Tomo achieves 0.60 AUROC – exceeding MedSigLIP w/ mean-pool (0.51) and aggregator (0.54) by 17.6% and 11.1% – confirming that aggregation preserves target-domain image–text alignment. Under LP, Mammo2Tomo yields an AUROC of 0.75 vs. 0.63 of the MedSigLIP w/ aggregator ($\sim$19%$\uparrow$). Contrasting our aggregator with mean pooling under each fixed encoder isolates the aggregator's contribution: mean pooling degrades performance via uniform slice averaging. Pretrained on non-mammographic medical imaging corpora, generalist MedSigLIP fails to capture breast-specific features for prototype matching. Mammo-CLIP-UnAdap. fails due to the 2D-to-3D domain shift, motivating our Stage 1 adaptation.

Data Efficiency. Fig. 2 shows that Mammo2Tomo outperforms the baselines for the data efficiency via LP across varying training subsets (25% to 100%). Notably, Mammo2Tomo trained on a 25% subset of the target data matches the cancer classification AUROC of the Mammo-CLIP-UnAdap. w/ aggregator ($\sim$0.72) and outperforms the fully trained MedSigLIP w/ aggregator ($\sim$0.63).

Text-guided slice retrieval. Tab. 3 shows Mammo2Tomo improves T2I slice retrieval over MedSigLIP and Mammo-CLIP-UnAdap., raising mean R@1 from 0.23 to 0.31 (35%$\uparrow$) and R@3 from 0.36 to 0.44 (22%$\uparrow$), with a 136% R@1 gain on Architectural Distortion (0.26 vs. 0.11). Qualitative MSAE localization (Fig. 3)

shows top-ranked slices activating on the ground-truth boxes, staying localized even for offset slices ($\Delta \neq 0$).

3 Conclusion

We introduce Mammo2Tomo, the first framework for text-aligned DBT volume representations. To overcome the lack of paired target data, Mammo2Tomo adapts Mammo-CLIP to mitigate the 2D-to-3D domain shift, aggregating slice embeddings into unified volumes via C-view-anchored masked reconstruction. Future work explores extending this framework to breast MRI.

Acknowledgments. This work was supported in part by the Pennsylvania Department of Health; NIH Award NIH 2R01HL141813-07; NSF Career Award NSF 2443167; the Hariri Institute for Computing at Boston University; and computational resources provided by the Pittsburgh Supercomputing Center (grant TG-ASC170024). Additional support was provided by NIH/NCI U01 CA269264-01-1.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Święcicki, A., Lo, J.Y., Mazurowski, M.A.: Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5,060 patients and a deep learning model. JAMA Network Open **4**(8), e2119100 (Aug 2021). https://doi.org/10.1001/jamanetworkopen.2021.19100
2. Chen, X., Li, Y., Hu, M., Salari, E., Chen, X., Qiu, R.L.J., Zheng, B., Yang, X.: Mammo-CLIP: Leveraging Contrastive Language-Image Pre-training (CLIP) for Enhanced Breast Cancer Diagnosis with Multi-view Mammography (Apr 2024). https://doi.org/10.48550/arXiv.2404.15946
3. Dorfner, F.J., Dorster, M.A., Connolly, R., Gentilhomme, O., Gibbs, E., Graham, S., Wander, S., Schultz, T., Bahl, M., Daye, D., Kim, A.E., Bridge, C.P.: DBT-DINO: Towards Foundation model based analysis of Digital Breast Tomosynthesis (Dec 2025). https://doi.org/10.48550/arXiv.2512.13608
4. Du, Y., Chen, L., Dvornek, N.C.: GLAM: Geometry-Guided Local Alignment for Multi-View VLP in Mammography (Sep 2025). https://doi.org/10.48550/arXiv.2509.10344
5. Du, Y., Onofrey, J., Dvornek, N.C.: Multi-View and Multi-Scale Alignment for Contrastive Language-Image Pre-training in Mammography (Mar 2025). https://doi.org/10.48550/arXiv.2409.18119
6. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-Adversarial Training of Neural Networks (May 2016). https://doi.org/10.48550/arXiv.1505.07818
7. Ghosh, S., Joshi, V.P., Syed, R., Kassem, A., Varshney, A., Basak, P., Dai, W., Gichoya, J.W., Trivedi, H.M., Banerjee, I., Visweswaran, S., Poynton, C.B., Batmanghelich, K.: Mammo-FM: Breast-specific foundational model for Integrated Mammographic Diagnosis, Prognosis, and Reporting (Nov 2025). https://doi.org/10.48550/arXiv.2512.00198

8. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography (May 2024). https://doi.org/10.48550/arXiv.2405.12255
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners (Dec 2021). https://doi.org/10.48550/arXiv.2111.06377
10. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilies, G., Smith, G., Woo, M., McAdams, C.R., Newell, M.S., Banerjee, I., Gichoya, J., Trivedi, H.: The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images. Radiology: Artificial Intelligence **5**(1), e220047 (2023). https://doi.org/10.1148/ryai.220047
11. Klein, D.S., Lago, M.A., Abbey, C.K., Eckstein, M.P.: A 2D Synthesized Image Improves the 3D Search for Foveated Visual Systems. IEEE transactions on medical imaging **42**(8), 2176–2188 (Aug 2023). https://doi.org/10.1109/TMI.2023.3246005
12. Lai, Z., Vesdapunt, N., Zhou, N., Wu, J., Huynh, C.P., Li, X., Fu, K.K., Chuah, C.N.: PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16109–16119. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.01480
13. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting Batch Normalization For Practical Domain Adaptation (Nov 2016). https://doi.org/10.48550/arXiv.1603.04779
14. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**(1), 277 (May 2023). https://doi.org/10.1038/s41597-023-02100-7
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision (Feb 2024). https://doi.org/10.48550/arXiv.2304.07193
16. Radiological Society of North America: RSNA Screening Mammography Breast Cancer Detection. https://www.kaggle.com/competitions/rsna-breast-cancer-detection (2022)
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
18. Sprague, B.L., Coley, R.Y., Lowry, K.P., Kerlikowske, K., Henderson, L.M., Su, Y.R., Lee, C.I., Onega, T., Bowles, E.J.A., Herschorn, S.D., diFlorio Alexander, R.M., Miglioretti, D.L.: Digital Breast Tomosynthesis versus Digital Mammography Screening Performance on Successive Screening Rounds from the Breast Cancer Surveillance Consortium. Radiology **307**(5), e223142 (May 2023). https://doi.org/10.1148/radiol.223142
19. Sun, B., Feng, J., Saenko, K.: Correlation Alignment for Unsupervised Domain Adaptation (Dec 2016). https://doi.org/10.48550/arXiv.1612.01939
20. Yang, Y., Soatto, S.: FDA: Fourier Domain Adaptation for Semantic Segmentation (Apr 2020). https://doi.org/10.48550/arXiv.2004.05498
21. Zaigrajew, V., Baniecki, H., Biecek, P.: Interpreting CLIP with Hierarchical Sparse Autoencoders (May 2025). https://doi.org/10.48550/arXiv.2502.20578

22. Zurrin, R., Goyal, N., Bendiksen, P., Manocha, M., Simovici, D., Haspel, N., Pomplun, M., Haehn, D.: Outlier detection for mammograms. In: International conference on medical imaging with deep learning (2023)