Geometry-Guided Local Alignment for Multi-View Visual Language Pre-Training in Mammography

Yuexi Du[1], Lihui Chen[1], Nicha C. Dvornek[1,2]

[1] Department of Biomedical Engineering,
[2] Department of Radiology & Biomedical Imaging,
Yale University, New Haven, CT, USA
{yuexi.du, leon.chen, nicha.dvornek}@yale.edu

Abstract. Mammography screening is an essential tool for early detection of breast cancer. The speed and accuracy of mammography interpretation has the potential to be improved with deep learning methods. However, the development of a foundation visual language model (VLM) is hindered by limited data and domain differences between natural and medical images. Existing mammography VLMs, adapted from natural images, often ignore domain-specific characteristics, such as multi-view relationships in mammography. Unlike radiologists who analyze both views together to process ipsilateral correspondence, current methods treat them as independent images or do not properly model the multi-view correspondence learning, losing critical geometric context and resulting in suboptimal prediction. We propose **GLAM**: **G**lobal and **L**ocal **A**lignment for **M**ulti-view mammography for VLM pretraining using geometry guidance. By leveraging the prior knowledge about the multi-view imaging process of mammograms, our model learns local cross-view alignments and fine-grained local features through joint global and local, visual-visual, and visual-language contrastive learning. Pretrained on EMBED [14], one of the largest open mammography datasets, our model outperforms baselines across multiple datasets under different settings. [3]

Keywords: Deep Learning · Visual-Language Pre-training · Contrastive Learning · Multi-view Alignment · Mammography

1 Introduction

Mammography screening is an effective tool for early detection of breast cancer, one of the most deadly cancers [26,28]. Unlike many natural or medical images that offer a single view, standard mammography protocol produces two 2D images of the same 3D breast from different angles – craniocaudal (CC) and mediolateral oblique (MLO) (Fig. 1(a)). This dual-view nature, known as *ipsilateral correspondence*, requires special consideration in clinical interpretation. Radiologists rely

[3] The code is available at https://github.com/XYPB/GLAM.

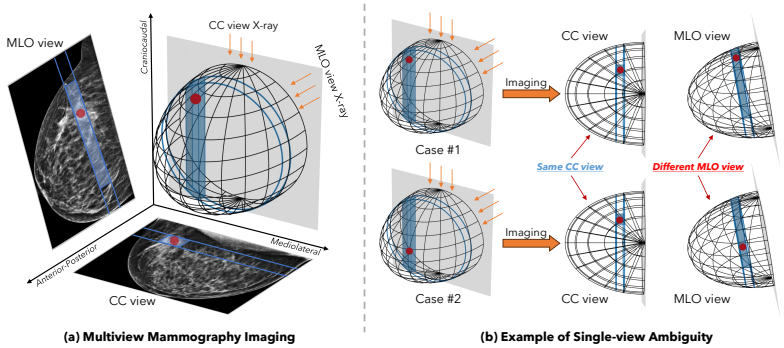

Fig. 1. The Importance of Multi-view. (a) Due to the imaging process, an ROI (red dot) appears in the same anterior-posterior (AP) slice in both mammography views. (b) Two ROIs located at different positions in the same AP slice result in the same CC view image but different MLO view images, demonstrating the single-view ambiguity.

on both views to accurately locate regions of interest (ROIs), such as tumors or calcifications, and to mitigate ambiguities caused by projection angles [13,18,15]. For instance, as in Fig. 1(b), an ROI (red dot) might lie anywhere along the vertical blue "tube", resulting in the same CC view image, while their MLO appearance will be different due to the MLO imaging angle. Thus, ignoring either view can lead to diagnostic errors, especially in data-driven deep-learning models that lack prior knowledge of the imaging process. In addition to considering prior imaging knowledge, inclusion of multi-modal information through contrastive language-image pre-training (CLIP) [24] has shown promise in enhancing medical image analysis. However, most prior CLIP models in the medical domain focus on other modalities like chest X-ray [29,35,30,31]. Meanwhile, mammography-specific models only conduct global alignment, neglecting fine-grained multi-view local alignment [5,12,10]. Besides, existing image-only multi-view mammography methods primarily use global feature fusion [1,6,32,27,16,20,7], which compromises local detail. Others consider local multi-view alignment, *e.g.*, using graph neural networks to learn the cross-view attention [18,17] or feature cosine similarity to model the multi-view relationship [13]; however, they lack the geometry knowledge needed for correct alignment that follows the actual 3D breast structure.

In this paper, we propose **G**lobal and **L**ocal **A**lignment for the **M**ulti-view mammography CLIP foundation model with geometry guidance, *i.e.*, **GLAM**. Pre-trained on $\sim 200k$ screening mammograms, our model is among the largest in this domain. Inspired by the mammography imaging process and geometry-guided patch matching [11,25], we propose a self-supervised, cross-view local patch alignment method that respects the CC and MLO projection relationship. Instead of patch-to-patch alignment that improperly treats the breast as a rigid body [18,33,13,19], we adopt patch-to-slice alignment along the anterior-posterior (AP) axis (Fig. 1(b)). We use cross-attention to include all relevant tissues along the AP slice while accounting for breast deformation in the CC-mediolateral (ML)

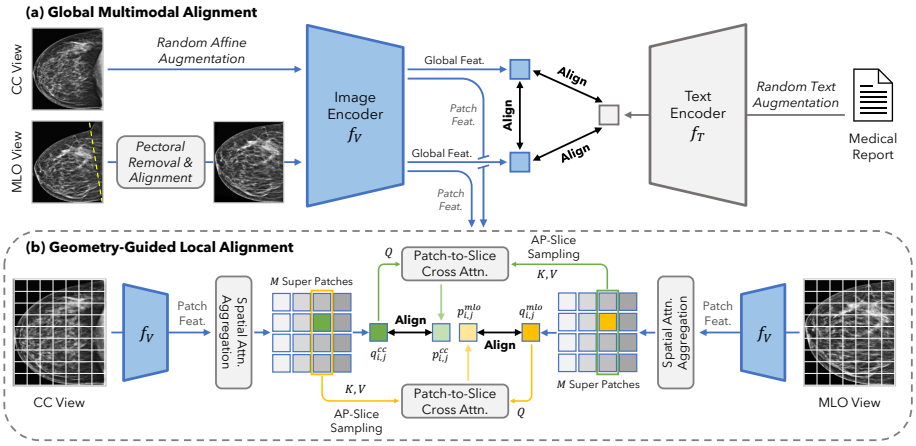

Fig. 2. GLAM Model. (a) Our method conducts global multi-view CLIP and aligns global visual features to text features from the report. (b) The patch level feature from each view is used to conduct geometry-guided local alignment, where patches from the same AP slice are used as positive matches in a cross-attention mechanism.

plane during imaging. We evaluate our method on three datasets with varied distributions [14,3,23], outperforming all baselines in multiple downstream tasks.

2 Methods

Given a multi-modal and multi-view mammography dataset $\mathcal{D} = \{(x_i^{cc}, x_i^{mlo}, y_i)\}$, where $i = 1, \ldots, N$, (x_i^{cc}, x_i^{mlo}) is the multi-view image pair and y_i is the radiology report. Our goal is to learn a robust mammography encoder f_V with both global multi-modal knowledge and local multi-view correspondence awareness (Fig. 2).

Pre-processing. Since the MLO view imaging is not parallel to the CC-ML plane (Fig. 1), the mammogram in the MLO view is inclined and contains a pectoral region. We remove the pectoral region using the Hough detector and rotate the image so that the segment between the chest and nipple is parallel to the AP axis, which better aligns the CC and MLO view along the AP axis. However, it is still possible that the two views are misaligned in the AP axis due to extreme cases such as a large pectoral region. We apply a random affine transformation to provide a soft alignment so that the model is more robust to local misalignment. Lastly, we synthesize the radiology report from tabular data following [10], which provides informative structured mammography reports, including imaging information, patient data, and findings. Random text augmentation following [34] is used to generate more diverse reports.

Contrastive Loss. We first define the contrastive loss $\mathcal{L}$ between two batched embeddings z and $\tilde{z}$ of size B in Eq. (1), following InfoNCE [4] loss:

$$\mathcal{L}(z, \tilde{z}) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\langle z_i,\ \tilde{z}_i \rangle / \tau)}{\sum_{j=1}^{B} \exp(\langle z_i,\ \tilde{z}_j \rangle / \tau)}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity and τ is the learnable temperature constant. All our learning objectives follow this basic contrastive form.

2.1 Global Multi-view Visual Language Pre-training

We first conduct multi-view visual language pre-training (VLP) at a global level. We extract the visual feature (v^{cc}, v^{mlo}) and textual feature t with corresponding modality encoder f_V and f_T. We use the embedding of `[CLS]` token as the global feature and optimize the multi-view contrastive loss $\mathcal{L}(v^{cc}, v^{mlo})$. Since the multi-view mammograms are different projections of the same breast, representing different views of the same information, it is natural to optimize the image-to-text contrastive loss symmetrically. Thus, our final global optimization objective is:

$$\mathcal{L}_{global} = \mathcal{L}(v^{cc}, v^{mlo}) + \frac{1}{2}\big[\mathcal{L}(v^{cc}, t) + \mathcal{L}(t, v^{cc}) + \mathcal{L}(v^{mlo}, t) + \mathcal{L}(t, v^{mlo})\big], \quad (2)$$

which optimizes both global multi-view contrastive loss and symmetric image-to-text losses. The textual supervision signal can help the model to learn an embedding space with high-level semantic information.

2.2 Geometry-Guided Local Alignment

Spatial Attention Aggregation. We use the patch features from f_V to conduct local alignment. Instead of using raw patch tokens that have a small receptive field, we aggregate the patch features using a spatial attention pooling layer to form M super-patches with a larger receptive field. These super-patches contain higher-level semantic information, which will be used for local alignment.

AP Slice Sampling and Local Alignment. We use the known geometry of mammography imaging as guidance to align the local patches. Namely, the image slices from both views in the same AP position represent the same tissue in the 3D breast, and each patch in the CC/MLO view should be aligned with a complete slice in the other view. Thus, we conduct patch-to-slice alignment along the AP axis. For a query patch $q^{cc}_{i,j}$ in i^{th} row and j^{th} column, its corresponding AP slice in the MLO view is $s^{mlo}_j = \{q^{mlo}_{1,j}, \ldots, q^{mlo}_{\sqrt{M},j}\}$. A multi-head cross-attention module is employed to model the alignment process between $q^{cc}_{i,j}$ and s^{mlo}_j,

$$p^{cc}_{i,j} = \text{CrossAttn.}\big(q^{cc}_{i,j},\ s^{mlo}_j,\ s^{mlo}_j\big) = \text{softmax}\big(\langle q^{cc}_{i,j},\ s^{mlo}_j\rangle/\sqrt{d}\big) \cdot s^{mlo}_j, \quad (3)$$

where d is the embedding dimension. We omit the linear projector for simplicity. The output $p^{cc}_{i,j}$ can be viewed as the weighted sum over the slice s^{mlo}_j based on its correspondence to the query patch. So, $p^{cc}_{i,j}$ will naturally be the cross-view positive for query patch $q^{cc}_{i,j}$. Similar computation is used for the MLO patches.

Negative Samples. To enhance the local positional awareness within the mammograms, we use all other patches from different positions as the negatives, *i.e.*, $\mathcal{S}^{cc}_{position} = \{p^{cc}_{m,n} | (m,n) \neq (i,j)\}$, which provides $M-1$ negative samples. However, using only $\mathcal{S}_{position}$ as negatives may result in a sub-optimal performance

Table 1. BI-RADS Prediction Results on EMBED. Performance (in %) for each method under zero-shot, linear probing with varying training data size, and full fine-tune settings. * denotes use of official pre-trained weights. Best and second-best results are in bold and underlined, respectively. Our method is shaded in gray.

Methods	Zero-shot		Linear Probing						Full Fine-tune	
	100%		1%		10%		100%		100%	
	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC
Vision only										
Random-ViT [9]	-	-	35.19	52.56	36.36	52.79	36.05	52.76	35.73	52.42
DiNOv2-ViT [22]	-	-	41.48	57.62	45.97	61.64	45.45	61.53	43.46	60.33
CLIP pre-trained										
CLIP [24]	37.17	55.90	43.35	59.57	47.89	64.46	47.05	63.50	45.77	61.79
SLIP [21]	44.24	60.67	43.43	60.39	48.82	64.67	46.66	63.35	37.81	54.60
ConVIRT [35]	43.02	61.31	47.45	63.16	47.91	63.78	47.73	63.40	49.41	65.41
MGCA [29]	45.48	61.92	47.81	62.76	48.30	63.44	48.82	64.83	50.37	65.70
Mammo-CLIP-B2* [12]	36.93	56.20	42.05	60.67	42.90	61.80	42.53	62.18	43.03	60.75
Mammo-CLIP-B5* [12]	36.67	57.09	38.68	59.93	38.15	61.21	38.29	61.25	38.58	61.46
MaMA [10]	44.61	61.63	46.63	63.65	48.90	64.81	47.96	63.69	49.95	66.06
GLAM (Ours)	**47.24**	**64.86**	**48.57**	**64.71**	**49.17**	**65.29**	**50.07**	**66.59**	**51.81**	**67.34**

as the model can learn to short-cut via using positional encoding. To address this, we use additional negative patches from the same position of different patients across the batch, *i.e.*, $\mathcal{S}^{cc}_{patient} = \{\tilde{p}^{cc}_{i,j} | \tilde{p}^{cc}_{i,j} \neq p^{cc}_{i,j}\}$ and $\tilde{p}^{cc}$ comes from other patients in the batch. These patches are natural negative samples for the query patch since they are from different patients; this forces the model to focus more on patch features rather than positional encoding. The final negative sample set is $\mathcal{S}^{cc} = \mathcal{S}^{cc}_{position} \cup \mathcal{S}^{cc}_{patient}$, providing $M + B - 2$ negative samples. The negative set for the MLO view query patches is built similarly.

Final Losses. We optimize the following local alignment loss symmetrically:

$$\mathcal{L}_{local} = -\frac{1}{2M} \sum_{i,j=1}^{\sqrt{M}} \sum_{\nu \in \{cc,mlo\}} \log \frac{\exp(\langle q^{\nu}_{i,j},\ p^{\nu}_{i,j} \rangle / \tau)}{\sum_{p^{\nu} \in \mathcal{S}^{\nu}} \exp(\langle q^{\nu}_{i,j},\ p^{\nu} \rangle / \tau)}. \tag{4}$$

$\mathcal{L}_{local}$ forces the model to align each super-patch with its corresponding AP slice from the other view and ensures the model learns both relative positional relationships and semantic correspondence across both views. The final optimization goal is the sum of global and local loss: $\mathcal{L}_{final} = \mathcal{L}_{global} + \mathcal{L}_{local}$.

3 Experiments

3.1 Experimental Settings

Datasets. We pre-train our model on the **EMBED** [14] dataset with over 257k screening mammograms with tabular annotated data. We create training/validation/test sets with 70%/10%/20% data, respectively. We evaluate on this dataset for screening BI-RADS (3 classes) and density (4 classes) prediction. We also evaluate on the **VinDr** [23] dataset with 20k images for BI-RADS (5 classes) and density (4 classes) prediction using the given data splits. This dataset

Table 2. Density Prediction Results on EMBED. Performance (in %) for each method under zero-shot, linear-probing with varying training data size, and full fine-tune settings. * denotes use of official pre-trained weights. Best and second-best results are in bold and underlined, respectively. Our method is shaded in gray.

Methods	Zero-shot		Linear Probing						Full Fine-tune	
	100%		1%		10%		100%		100%	
	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC
Vision only										
Random-ViT [9]	-	-	38.99	68.99	41.35	68.91	41.48	69.03	64.55	86.44
DiNOv2-ViT [22]	-	-	65.62	87.06	67.54	87.39	67.54	87.45	77.47	93.40
CLIP pre-trained										
CLIP [24]	59.69	88.73	74.77	<u>93.32</u>	76.15	91.83	76.92	92.91	78.32	<u>93.91</u>
SLIP [21]	<u>78.06</u>	<u>92.78</u>	75.47	93.19	<u>77.22</u>	92.90	<u>77.99</u>	<u>93.75</u>	<u>78.81</u>	93.89
ConVIRT [35]	61.48	72.54	73.64	92.77	74.20	92.16	75.27	92.70	78.31	93.85
MGCA [29]	62.45	71.29	72.34	91.48	72.89	91.92	73.56	91.98	78.37	93.66
Mammo-CLIP-B2* [12]	53.50	80.50	70.02	88.91	68.98	88.59	69.22	88.81	76.01	92.47
Mammo-CLIP-B5* [12]	46.07	71.89	69.60	89.47	70.23	89.98	69.46	89.96	69.90	90.05
MaMA [10]	75.18	91.81	<u>74.88</u>	92.79	76.74	<u>93.15</u>	73.67	91.69	77.61	92.66
GLAM (Ours)	**79.06**	**93.76**	**77.87**	**93.65**	**78.76**	**94.01**	**79.61**	**94.03**	**80.32**	**94.05**

(from Vietnam) has a different distribution than our pre-training data (from USA). We also evaluate on the **RSNA-Mammo** [3] dataset for binary cancer prediction. From the provided dataset of 54k images, we split 15% as the test set.

Tasks. We focus on three classification settings for individual mammograms: *zero-shot* on in-domain data, *linear-probing*, and *full fine-tune*. We further vary the size of training data in linear probing to evaluate the data efficiency.

Implementations. We initialize our encoders using BioClinical-BERT [2] and DiNOv2 [22] ViT-B [9]. We use a batch size of 144, learning rate of 4×10^{-5}, and weight decay of 0.2 to pre-train our model using SGD and cosine learning rate scheduler for 40k steps. For downstream linear probing and fine-tuning, we set batch size to 96, learning rate to 5×10^{-4}, and weight decay to 0.001 and train for 8k steps using SGD. We train with balanced sampling. The same setting is applied to all baselines. All images are resized to 518×518 as input.

Baselines and Metrics. We compare with vision-only transformers with or without ImageNet [8] pre-training; CLIP [24] and SLIP [21] as natural image domain baselines; ConVIRT [35] and MGCA [29] as medical CLIP baselines from different imaging domains; and Mammo-CLIP [12] and MaMA [10] as in-domain baselines. *All baselines except Mammo-CLIP are pre-trained on EMBED* [14], just like our model. We use the official pre-trained weights for Mammo-CLIP to show the influence of different pre-training data. We report balanced accuracy (bACC) and AUC as our metrics since the distribution of mammography data is extremely imbalanced; simple accuracy may be biased toward majority classes.

3.2 Results

In Domain Analysis. We first evaluate on the in-domain EMBED test set for BI-RADS and density prediction (Tab. 1 and Tab. 2). Our model outperforms all the baselines consistently in all scenarios, surpassing the best baselines by 2.3% in AUC on average. Even with only 1% of training data, our pre-trained

Table 3. Results on VinDr and RSNA-Mammo. Performance (in %) for each method on prediction tasks on VinDr and RSNA-Mammo datasets under linear-probing and full fine-tune settings. * denotes use of official pre-trained weights. Best and second-best results are in bold and underlined, respectively. Our method is shaded in gray.

Methods	VinDr - BI-RADS				VinDr - Density				RSNA - Cancer			
	Linear Probing		Full Fine-tune		Linear Probing		Full Fine-tune		Linear Probing		Full Fine-tune	
	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC
Vision only												
Random-ViT [9]	27.28	56.50	26.58	59.62	29.69	56.95	34.25	37.46	54.75	57.06	54.18	58.31
DiNOv2-ViT [22]	35.47	63.51	37.99	64.36	59.68	86.38	69.49	92.96	52.94	61.01	52.25	66.57
CLIP pre-trained												
CLIP [24]	41.26	68.92	41.11	72.36	70.85	93.07	72.38	92.52	65.79	71.93	63.69	67.94
SLIP [21]	40.03	70.18	41.06	74.23	71.98	92.79	66.37	85.80	60.05	65.78	55.65	61.14
ConVIRT [35]	39.11	71.64	39.95	72.18	63.63	71.01	71.01	90.05	62.74	68.80	53.59	66.03
MGCA [29]	38.85	72.72	40.85	73.27	71.82	89.63	76.94	90.51	64.64	69.94	54.69	68.46
Mammo-CLIP-B2* [12]	34.68	64.76	36.23	65.92	64.09	87.90	64.26	87.91	52.81	61.52	53.55	61.32
Mammo-CLIP-B5* [12]	39.68	67.58	**42.78**	71.83	70.70	87.64	**78.56**	93.28	60.72	66.02	64.50	72.96
MaMA [10]	41.35	68.36	35.94	61.78	73.49	92.77	65.63	92.64	63.18	69.32	57.31	62.28
GLAM (Ours)	**41.41**	**73.81**	41.87	**74.82**	**74.58**	**93.60**	78.27	**93.94**	**67.45**	**73.14**	**68.77**	**75.04**

Table 4. Multi-view Prediction Results. Zero-shot performance (in %) of BI-RADS and density prediction under single-view and multi-view settings. * denotes use of official pre-trained weights. Best results are in bold. Our method is shaded in gray.

Methods	EMBED - BI-RADS				EMBED - Density			
	Single-view		Multi-view		Single-view		Multi-view	
	bACC	AUC	bACC	AUC	bACC	AUC	bACC	AUC
CLIP [24]	42.35	61.79	44.47	63.56	57.22	87.40	57.65	88.08
MGCA [29]	44.72	62.17	46.17	63.23	68.49	89.87	71.23	91.30
Mammo-CLIP-B2* [12]	36.80	56.70	36.36	56.98	53.29	80.24	54.65	81.04
Mammo-CLIP-B5* [12]	38.27	58.35	38.30	58.64	49.63	72.63	50.20	73.42
GLAM (Ours)	**46.05**	**63.28**	**48.40**	**66.02**	**79.02**	**93.63**	**79.42**	**94.08**

model can still outperform almost all baselines trained with 100% of data. Vision-only methods generally underperform models with VLP. We note that Mammo-CLIP [12], pre-trained on ∼10× less data, is 8% lower on average in bACC, showing a worse generalization capability and highlighting the necessity of scaling the training data. Meanwhile, other baselines [10,29,21] that have only global multi-view alignment failed to beat our model since they lack fine-grained multi-view awareness, resulting in suboptimal embedding space.

Out of Domain Analysis. We further evaluate performance on the out-of-domain datasets VinDr and RSNA-Mammo (Tab. 3) to illustrate the generalization capability of each model. Our model performs the best in 10 out of 12 metrics, suggesting good generalization on unseen data. We note that the gap between our method and other baselines is smaller in full fine-tune settings compared with linear probing. This is mainly because these out-of-domain datasets have a smaller training set, which makes it easier for the model to converge.

Multi-view Analysis. We evaluate the capability of modeling multi-view correspondence under zero-shot settings, which focus on pre-trained embedding quality (Tab. 4). We sub-sampled a test set from EMBED with 7,676 paired multi-view mammograms and evaluated under single- and multi-view prediction, where the multi-view prediction is obtained by averaging the single-view results. Our model improves by ∼2.5% in BI-RADS prediction after switching to multi-view settings,

Table 5. Ablation Results. Performance (in %) of BI-RADS prediction on EMBED for each ablated model. GLA: Geometry-guided Local Alignment; SPN: Same Position Negatives; SAA: Spatial Attention Aggregation. Best results are in bold. Our method is shaded in gray.

GLA	SPN	SAA	AP Sampling	#Local Regions			Zero-shot		Linear Probing		Full Fine-tune	
				$M = 16$	$M = 81$	$M = 324$	bACC	AUC	bACC	AUC	bACC	AUC
							45.12	62.82	49.36	65.33	37.81	54.60
✓		✓	✓		✓		45.55	62.23	47.48	63.75	36.34	53.85
✓	✓		✓		✓		44.24	60.67	46.66	63.35	37.68	53.99
✓	✓	✓			✓		44.98	62.56	49.17	65.24	48.45	64.45
✓	✓	✓	✓	✓			43.75	60.81	46.94	64.03	46.80	63.51
✓	✓	✓	✓			✓	45.99	62.72	48.55	64.76	47.58	63.07
✓	✓	✓	✓		✓		**47.24**	**64.86**	**50.07**	**66.59**	**51.81**	**67.34**

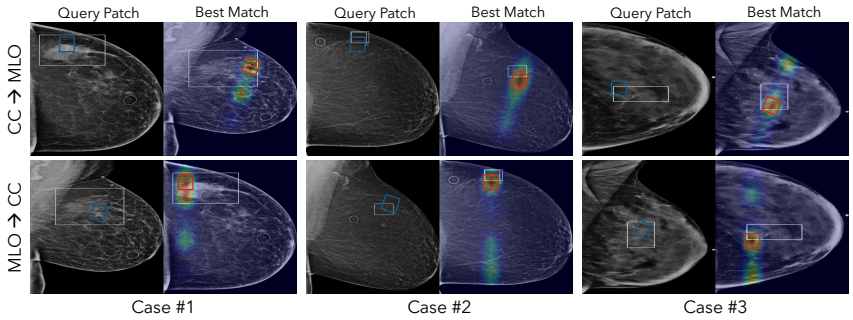

Fig. 3. Cross View Patch-to-Slice Attention Visualization For each pair of mammograms, the white bounding boxes indicate the ROIs, *e.g.*, tumor; the blue box is the query patch; and the red box is the patch with the highest attention. Patches in the MLO view are inclined due to AP alignment during pre-processing.

while the baselines have less to no improvement. Since the density is similar in both views, there is less improvement. This indicates that our method can model the multi-view geometry and extract complementary features for each view.

Ablation Study. Model ablation results are in Tab. 5. First, removing the geometry-guided local alignment learning greatly harms the model's performance, especially its robustness under full fine-tuning. Same-position negatives provide $\sim$2% improvement in zero-shot and linear probing and ensure stable behavior in full fine-tuning. Replacing the spatial attention aggregation with average pooling also results in sub-optimal performance. Lastly, we evaluate the necessity of following the geometry guidance in local correspondence learning by computing the attention across all patches in the view. This lowers performance since the geometry constraint is broken. We also test different numbers of super-patches M, where patch size will influence the performance as discussed in Sec. 2.2.

Qualitative Visualization. We visualize the cross-view patch-to-slice attention weights in Fig. 3. We pick random patches within annotated ROIs from the EMBED test set and visualize their attention scores in the corresponding AP slice in the other view. Our model can accurately locate the ROI in the other view and, therefore, gain multi-view awareness during pre-training.

4 Discussion and Conclusion

We proposed one of the largest screening mammography foundation CLIP models to date, *i.e.*, GLAM, with a novel geometry-guided local alignment module to enable the fine-grained cross-view awareness of the model. The proposed method achieved state-of-the-art performance in three different datasets compared with existing VLP models. While we mainly focus on evaluating the quality of the pre-trained embedding space, we also plan to fuse our robust backbone with multi-view fusion methods to further improve the performance and clinical applicability. Future plans include introducing dense multi-modal contrastive learning and extending multi-view alignment to both sides of the breast.

Acknowledgments This work was supported by NIH grant R21EB032950.

Disclosure of Interests The authors have no competing interests in this work and other related research.

References

1. Akselrod-Ballin, A., et al.: Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology **292**(2), 331–342 (2019)
2. Alsentzer, E., McDermott, M., et al.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/W19-1909, https://www.aclweb.org/anthology/W19-1909
3. Carr, C., et.al., Y.C.: Rsna screening mammography breast cancer detection (2022), https://kaggle.com/competitions/rsna-breast-cancer-detection
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chen, X., Yang, X., et al.: Mammo-clip: Leveraging contrastive language-image pre-training (clip) for enhanced breast cancer diagnosis with multi-view mammography. arXiv preprint arXiv:2404.15946 (2024)
6. Chen, Y., Carneiro, G., et al.: Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 3–13. Springer (2022)
7. Chen, Y., Liu, Y., Wang, C., Elliott, M., Kwok, C.F., Peña-Solorzano, C., Tian, Y., Liu, F., Frazer, H., McCarthy, D.J., et al.: Braixdet: Learning to detect malignant breast lesion with incomplete annotations. Medical image analysis **96**, 103192 (2024)
8. Deng, J., Fei-Fei, L., et al.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on CVPR. pp. 248–255. Ieee (2009)
9. Dosovitskiy, A., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

10. Du, Y., Onofrey, J., Dvornek, N.C.: Multi-view and multi-scale alignment for contrastive language-image pre-training in mammography. arXiv preprint arXiv:2409.18119 (2024)
11. Engeland, S.v., Timp, S., Karssemeijer, N.: Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views. Medical Physics **33**(9), 3203–3212 (2006)
12. Ghosh, S., Batmanghelich, K., et al.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. arXiv preprint arXiv:2405.12255 (2024)
13. Jain, K., Rangarajan, K., Arora, C.: Follow the radiologist: Clinically relevant multi-view cues for breast cancer detection from mammograms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 102–112. Springer (2024)
14. Jeong, J.J., Smith, G., et al.: 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)
15. Ji, C., Shen, D., et al.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: MICCAI. pp. 68–78. Springer (2023)
16. Jouirou, A., Baâzaoui, A., Barhoumi, W.: Multi-view information fusion in mammograms: A comprehensive overview. Information Fusion **52**, 308–321 (2019)
17. Liu, Y., Yu, Y., et al.: Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: Proceedings of the IEEE/CVF conference on CVPR. pp. 3812–3822 (2020)
18. Liu, Y., Yu, Y., et al.: Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5947–5961 (2021)
19. Ma, J., Li, X., Li, H., Wang, R., Menze, B., Zheng, W.S.: Cross-view relation networks for mammogram mass detection. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8632–8638. IEEE (2021)
20. Manigrasso, F., Morra, L., et al.: Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures. Medical Image Analysis **99**, 103320 (2025)
21. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European conference on computer vision. pp. 529–544. Springer (2022)
22. Oquab, M., Bojanowski, P., et al.: Dinov2: Learning robust visual features without supervision (2023)
23. Pham, H.H., Trung, H.N., Nguyen, H.Q.: Vindr-mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography. Physionet https://doi. org/10.13026/br2v-7517 (2022)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Sahiner, B., Zhou, C., et al.: Joint two-view information for computerized detection of microcalcifications on mammograms. Medical physics **33**(7Part1), 2574–2585 (2006)
26. Siegel, R., Ma, J., Zou, Z., Jemal, A.: Cancer statistics, 2014. CA: a cancer journal for clinicians **64**(1) (2014)

27. Sun, Z., Jiang, H., Ma, L., Yu, Z., Xu, H.: Transformer based multi-view network for mammographic image classification. In: MICCAI. pp. 46–54. Springer (2022)
28. Sung, H., Bray, F., et al.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
29. Wang, F., Yu, L., et al.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems **35**, 33536–33549 (2022)
30. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
31. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. medRxiv pp. 2023–01 (2023)
32. Xia, L., Gao, Z., et al.: Neural network model based on global and local features for multi-view mammogram classification. Neurocomputing **536**, 21–29 (2023)
33. Yang, Z., Huang, L., et al.: Momminet-v2: Mammographic multi-view mass identification networks. Medical Image Analysis **73**, 102204 (2021)
34. You, K., Roh, B., et al.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
35. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)