MM-CLIP:Multi-granularity Mammography CLIP for Breast Cancer Diagnosis

Han Wei
Ocean Univisity Of China
Qingdao, Shandong, China
weihan@stu.ouc.edu.cn

Hao Fan
Ocean Univisity Of China
Qingdao, Shandong, China
fanhao@ouc.edu.cn

Abstract

Breast cancer is one of the most common cancers among women, and analyzing mammograms to assist in diagnosis is of great significance in the fields of computer vision and artificial intelligence. However, unlike natural image recognition, mammogram analysis requires focusing on small and complex lesion areas. The diversity of lesion morphology and the individual variability of anatomical structures further increase the complexity and challenges of diagnosis. This paper proposes a multi-granularity knowledge-guided multimodal pre-training method for breast cancer diagnosis. Experiments demonstrate that, compared to previous CLIP-based methods, the proposed approach is more effective.

CCS Concepts

• **Computing methodologies** → **Image processing**; *Multimodal reasoning*; *Machine learning*; *Neural networks*; • **Applied computing** → **Life and medical sciences**; *Health informatics*.

Keywords

Medical diagnostic methods,Medical Imgae,CLIP

ACM Reference Format:
Han Wei and Hao Fan. 2025. MM-CLIP:Multi-granularity Mammography CLIP for Breast Cancer Diagnosis. In *2025 9th International Conference on Digital Signal Processing (ICDSP 2025), February 21–23, 2025, Chengdu, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3725988.3726002

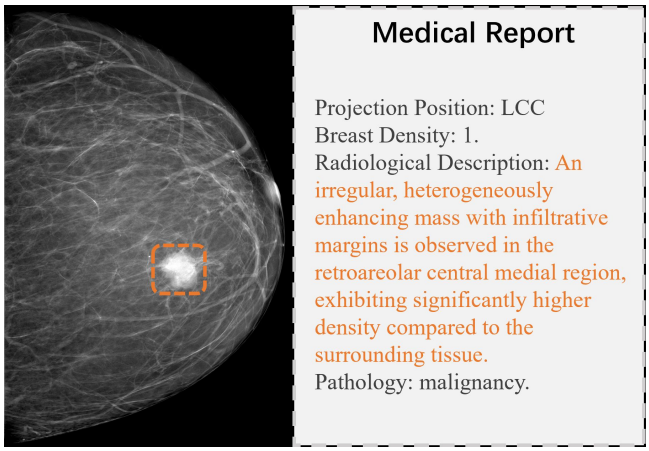

Figure 1: the semantic gap between image features and medical reports

1 Introduction

Cancer is one of the most threatening diseases today, and breast cancer is the most common cancer among women worldwide. According to statistics, breast cancer is listed as the most common cancer in women in 157 out of 185 countries globally[11], with over 2 million new cases diagnosed each year. Mammograms, as a crucial basis for diagnosing breast cancer, play a vital role in early detection and treatment. Meanwhile, advancements in artificial intelligence, such as Contrastive Language-Image Pretraining (CLIP), are revolutionizing medical imaging and diagnostics. CLIP, which establishes a bridge between images and texts through contrastive training on large-scale image-text pairs, has demonstrated strong transferability and generality across various downstream tasks, including zero-shot image classification and image-text retrieval. Its outstanding prior knowledge makes it a promising foundational framework for vision-language models, potentially enhancing the accuracy and efficiency of breast cancer diagnosis and other medical imaging applications.

In mammographic imaging, CLIP struggles to replicate its success with natural images. Breast cancer diagnosis relies on lesion features like masses, calcifications, and asymmetries, but global image labels fail to capture these local details, hindering accurate lesion analysis. While some methods align images with medical reports, challenges persist: lesion areas occupy only 2%-3% of the image, while reports focus heavily on these small regions. This creates a semantic gap between image features and report descriptions, making precise alignment a critical challenge.

To enhance the semantic association between lesion areas and medical reports, this work proposes a Multi-granularity Mammography Contrastive Image-Text Pretraining Network (MM-CLIP). To our knowledge, MM-CLIP is the first model to address the semantic gap between mammograms and medical reports using a region-level alignment strategy. The network employs multimodal contrastive learning, leveraging global image features and region of interest (ROI) features annotated by physicians to enhance semantic representation, achieving more accurate image-text alignment. During pretraining, the model uses contrastive learning to acquire discriminative features, mitigating imbalanced data distribution. Experiments show that MM-CLIP exhibits strong zero-shot recognition and achieves significant improvements in tasks like benign and

ICDSP 2025, Chengdu, China

ACM ISBN 979-8-4007-1046-9/25/02
https://doi.org/10.1145/3725988.3726002

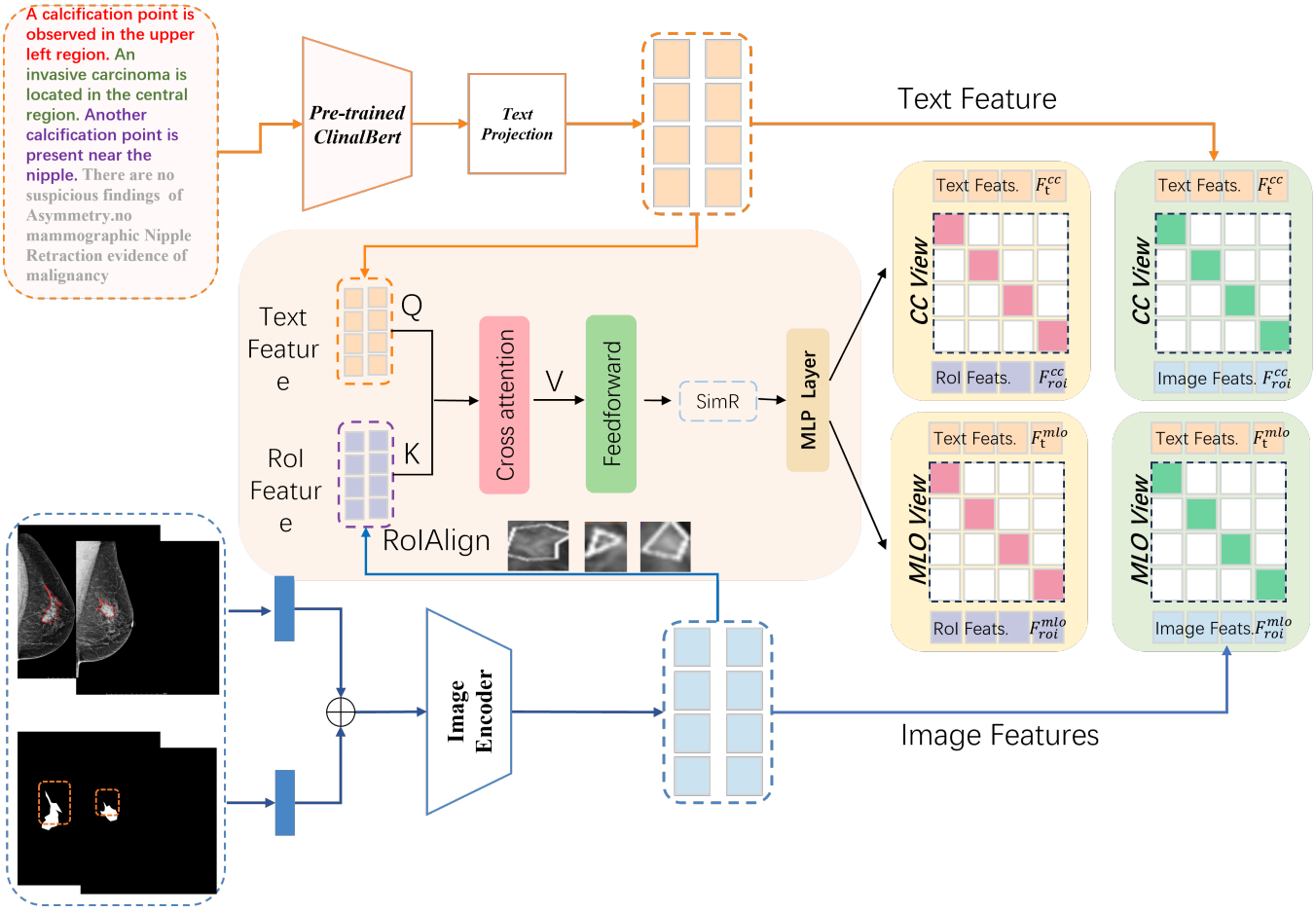

Figure 2: The structure of the Multi-granularity Mammography CLIP (MM-CLIP)

malignant classification of breast masses. The main contributions are as follows:

- We proposed a region-focused image feature generation module , which effectively incorporates prior knowledge of lesion areas in mammograms to achieve comprehensive representation of both lesion-specific and global image features.
- We introduced a feature matching module based on multi-head attention mechanisms, which aim to bridge the semantic gap between image features and medical reports.
- We proposed MM-CLIP, a novel CLIP pre-training paradigm, and validate it through experiments on multiple public datasets. Compared to existing methods, our approach demonstrates greater effectiveness in breast cancer diagnosis.

2 Related Work

Breast Canner Diagnosis: With the rapid development of deep learning, breast cancer classification algorithms based on mammography images have made significant progress in diagnosis. Zhu et al.[21] transformed the overall classification problem of mammograms into a Multiple Instance Learning (MIL) task by designing three MIL loss functions (max-pooling loss, label assignment loss, and sparse loss) to fine-tune the pre-trained AlexNet model. Zhang et al.[19] compared the performance of models such as ResNet50 based on transfer learning in the overall classification task. Shen et al.[9] utilized data annotated with regions of interest (ROI), integrating clinical annotations and overall image labels to enhance breast cancer classification performance. Shu et al.[10] proposed region group max-pooling (RGP) and global group max-pooling (GGP) methods based on the observation that cancerous tissues typically account for only 2%-3% of breast tissue. Addressing the challenge of medical image resolution, Ahamed et al.[12] introduced an efficient HCT model based on linear self-attention mechanisms to tackle long-range dependency issues. Tao et al.[16] designed a domain-specific network front-end to reduce the memory requirements of high-resolution images and improve model applicability. Han et al.[5] employed a position embedding module to extract positional information from two different dimensions of feature maps and converted it into weight maps to weight the original features; based on this, they used two pooling strategies to capture positional features at two scales, thereby selecting the most suspicious lesion regions.Petrini et al.[7] proposed a Cross-View Relation

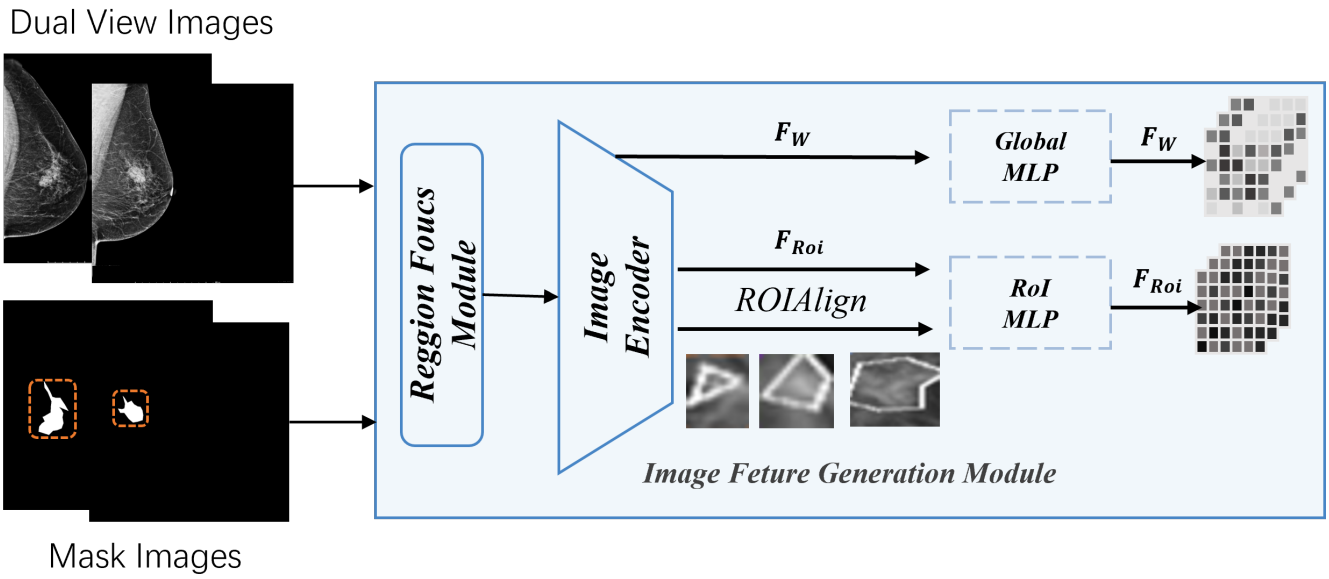

Figure 3: Image Feature generation Module

Region Convolutional Neural Network. Chen et al.[2] introduced a multi-view global-local analysis method, combining local and global information to further enhance the accuracy of mammogram classification. You et al.[18] incorporated contrastive learning, utilizing contrastive learning and triplet loss on normal and lesion samples to improve the separability of the embedding space. Wang et al.[14] enhanced model performance by introducing a dual-view correlation loss.

Medical CLIP: In medical imaging, Contrastive Language–Image Pretraining (CLIP) models can generally be categorized into two types based on their application scope: general-purpose medical CLIP models and domain-specific medical CLIP models.General-purpose medical CLIP models are typically pre-trained on large-scale medical image datasets, covering multiple anatomical regions and imaging modalities. These models focus on expanding dataset scale while using the original CLIP architecture [8]. ConVIRT [20] pioneered medical vision-language pre-training through contrastive learning, showcasing strong transfer learning with large-scale medical images and reports. BioViL [1] improved joint representation learning by aligning multiple image-report pairs, enhancing image-text semantic alignment. MedCLIP [15] integrated medical knowledge, replacing InfoNCE loss with a knowledge-based loss, boosting performance on tasks like chest X-rays and fundus images. PubMedCLIP [3] expanded datasets from PubMed, improving generalization. However, without modality-specific optimizations, these models underperform compared to specialized ones.

Domain-specific medical CLIP models have gained traction in mammography but face data limitations. Frozen [13] introduced CLIP to mammography by freezing the visual encoder and training only the classification head, skipping contrastive pre-training and ignoring pixel-level imbalances. Kshitiz [6] used CNNs to extract regions of interest for contrastive learning but overlooked global context. Ghosh et al. [4] proposed Mammo-CLIP but missed multi-scale characteristics, hindering precise lesion-report alignment. Current mammography CLIP models often focus on either global or local regions, lacking a multi-granularity approach. This limitation restricts their ability to fully capture the intricate interplay between macroscopic tissue patterns and microscopic lesion details, both of which are essential for accurate breast cancer diagnosis.

3 Method

As previously discussed, the core objective of the MM-CLIP model is to bridge the gap between the textual features of medical reports and the perceptual representations of lesion areas, thereby introducing high-level human prior knowledge to drive the feature encoding network of limited scale to extract richer semantic information. Specifically, Figure 2 illustrates the detailed architecture of MM-CLIP. Section 3.1 introduces the generation of image features, Section 3.2 describes the generation of text features, Section 3.3 explains the process of feature matching, and Section 3.4 presents the loss functions used in this study.

3.1 Image Feature generation Module

To simultaneously obtain feature representations of both the images and the regions of interest (RoIs) while preserving spatial information, as shown in Figure 3, this paper designs a multi-scale image feature generation module. The process within this module is described as follows: Given the original image set $I = (I^{cc}, I^{mlo})$ and the RoI annotation set $X = (X^{cc}, X^{mlo})$, where $X^{cc}, X^{mlo} \in n$ categories and $I^{cc}, I^{mlo} \in \mathbb{R}^{C\times H\times W}$. First, to enhance the feature representation of the RoIs in the original images, this paper generates corresponding masks $M^{cc}, M^{mlo} \in \mathbb{R}^{C\times H\times W}$ based on X^{cc}, X^{mlo}, and inputs the masks along with the original images into the region

attention module. Specifically, as shown in Figure 4, both are passed through a convolutional layer with a kernel size of 1×1, followed by element-wise addition, to capture shallow feature representations $F_l^{cc}, F_l^{mlo} \in \mathbb{R}^{C\times H\times W}$ that are richer in semantic information of the lesion regions.

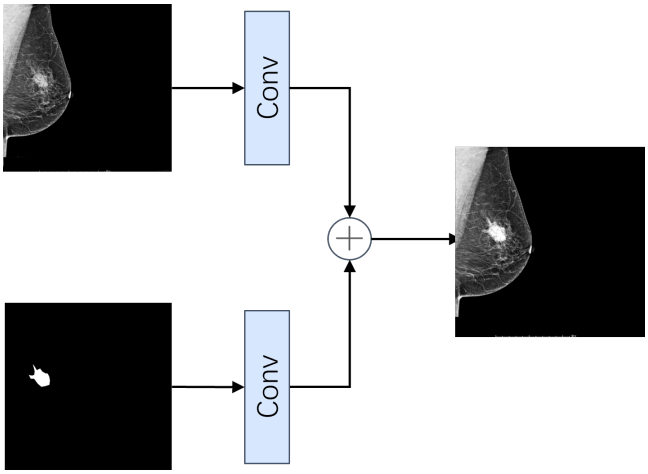

Figure 4: Region Focus Module

Furthermore, the obtained feature representations F_l^{cc}, F_l^{mlo} are fed into a visual encoder to acquire higher-level feature representations F_h^{cc}, F_h^{mlo}. These high-level features can model the complex morphology and structure of lesions, capturing intricate nonlinear relationships, thereby learning more discriminative features. Subsequently, this paper employs the RoIAlign layer to obtain high-level feature representations $F_{roi}^{cc}, F_{roi}^{mlo}$ corresponding to the lesion regions while preserving spatial information. Finally, this paper introduces an MLP-based global prediction layer and an RoI prediction layer to obtain the final feature representations $F_W = (F^{cc}, F^{mlo})$ and $F_{roi} = (F_{roi}^{cc}, F_{roi}^{mlo})$. The entire process is formally defined by the following Equations:

$$F_W = \begin{cases} F_{cc} = \Phi(\xi(I_{cc}, M_{cc})), & M_{cc} = \text{Mask}(X_{cc}); \\ F_{mlo} = \Phi(\xi(I_{mlo}, M_{mlo})), & M_{mlo} = \text{Mask}(X_{mlo}); \end{cases} \tag{1}$$

$$F_{roi} = \begin{cases} F_{roi}^{cc} = \psi(F_{cc}); \\ F_{roi}^{mlo} = \psi(F_{mlo}); \end{cases} \tag{2}$$

where ξ represents the region attention module processing the input images and masks; Mask denotes the conversion of RoI annotations into corresponding mask images; Φ represents the visual encoder extracting deep-level features; and ψ represents the RoIAlign layer obtaining features with spatial information and deep semantic information.

3.2 Text Feature generation Module

Due to the absence of prior textual descriptions related to mammographic images, we employ a structured report generation method based on Large Language Models (LLMs). As illustrated in Figure 5, this paper inputs bilateral image data into a visual encoder to extract preliminary textual descriptions. The visual encoder utilized is MedCLIP[15], a medical image-text comprehension model. Subsequently, this paper combines the preliminary textual descriptions with pixel-level and image-level annotations as prompts to be fed into a large language model, thereby obtaining structured medical reports. The LLM used here is GPT-4. The method presented in this paper can also be applied to common datasets such as Vindr, further expanding the usability of multimodal data. After generating the medical report,we input these reports as textual prompts into MM-CLIP.The specific steps are as follows: input text prompts W_{cc} and W_{mlo} are converted into tokens T_{cc} and T_{mlo} using a medical-optimized tokenizer, which better handles medical terminology. These tokens are processed by the pre-trained Bio_ClinicalBERT model to generate initial text features F_{cc} and F_{mlo}. A shared-weight deep neural network further refines these features to capture lesion semantics. To align image and text features, a multi-layer perceptron (MLP) maps both modalities into a unified space, enabling semantic consistency matching. This step is formally defined by the follow Equation:

$$F_t = \begin{cases} F_{cc} = \Phi(F(T_{cc})), & T_{cc} = \text{Token}(W_{cc}); \\ F_{mlo} = \Phi(F(T_{mlo})), & T_{mlo} = \text{Token}(W_{mlo}); \end{cases} \tag{3}$$

where Φ represents the text feature encoder based on the multilayer perceptron; F denotes the pre-trained Bio_ClinicalBERT model; and Token represents the processing of the original medical reports using the tokenizer.

3.3 Feature Matching Module

In CLIP-related research, the feature matching module processes image content and text descriptions using specialized encoders to extract image features F_w, F_{roi}, and text features F_t. A contrastive learning strategy is employed to project these features into a shared embedding space, enhancing their discriminative power and quantifying semantic similarity.

For a batch size of B, feature matching matrices M_W and M_{roi}, each of size $2\times B\times B$, are constructed. Positive samples are matched "image-text" pairs, while unmatched pairs are negative samples. The diagonal elements of the matching matrix correspond to the B positive sample pairs, representing their matching degree. The feature matching matrix for whole images is defined as:

$$M_w = \begin{cases} M_{cc} & = \alpha_0 \tilde{F}_{cc}^w \otimes (\tilde{F}_{cc}^t)^T, \\ M_{mlo} & = \alpha_1 \tilde{F}_{mlo}^w \otimes (\tilde{F}_{mlo}^t)^T. \end{cases} \tag{4}$$

Here, M_{cc} and M_{mlo} represent the similarity between dual-view images and text for the CC and MLO views, respectively. α_0 and α_1 are learnable temperature parameters, $\tilde{F}_*$ denotes normalized features, $\otimes$ represents matrix multiplication, and T denotes transposition.

A cross-modal attention module explores the relationship between regional image features F_{roi} and text features F_t. The image features serve as the query matrix (Q), and text features as the key matrix (K). The feature matching matrix is generated as:

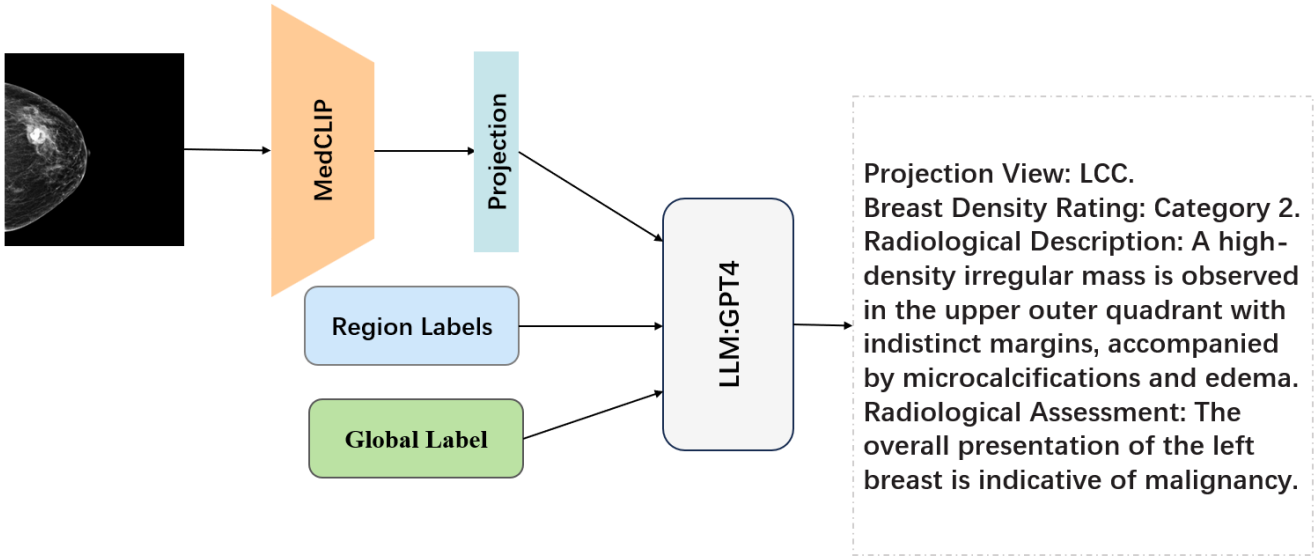

Figure 5: the pipeline of LLM report generation

$$M_{roi} = \begin{cases} M_{roi_{cc}} = \alpha_2 \mathcal{P}(\mathcal{T}(\mathcal{L}(Q_{cc}, K_{cc}))), \\ M_{roi_{mlo}} = \alpha_3 \mathcal{P}(\mathcal{T}(\mathcal{L}(Q_{mlo}, K_{mlo}))). \end{cases} \tag{5}$$

Here, $M_{roi_{cc}}$ and $M_{roi_{mlo}}$ are the feature matching matrices for dual-view lesion images and text; $\mathcal{P}$ denotes the MLP, $\mathcal{T}$ calculates similarity via multi-head attention, α_2 and α_3 are learnable parameters, and $\mathcal{L}$ unifies dimensions for fusion and similarity computation.

3.4 Loss Functions

This paper employs three loss functions to train the MM-CLIP model: cross-view consistency loss, global CLIP loss, and RoI CLIP loss. The basic CLIP loss is defined to maximize the diagonal values of the feature matching matrix, ensuring consistent matching between text and image features. The formula is:

$$\mathcal{L}_{\text{CLIP}_*} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log \frac{\exp(\hat{F}_{k^*} \cdot \hat{F}_{t_k^*}/\tau_*)}{\sum_{j=1}^{B}\exp(\hat{F}_{k^*} \cdot \hat{F}_{t_j^*}/\tau_*)} + \log \frac{\exp(\hat{F}_{t_k^*} \cdot \hat{F}_{k^*}/\tau_*)}{\sum_{j=1}^{B}\exp(\hat{F}_{t_k^*} \cdot \hat{F}_{j^*}/\tau_*)}\right] \tag{6}$$

Here, $\hat{F}_{k^*}$ and $\hat{F}_{t_k^*}$ are image and text embeddings, and τ_* is the temperature coefficient.

- **Global CLIP Loss**: Aligns overall semantic relationships between whole image features and text features. Defined as:

$$\mathcal{L}_{\text{CLIP}_W} = \frac{\mathcal{L}_{\text{CLIP}_{cc}} + \mathcal{L}_{\text{CLIP}_{mlo}}}{2} \tag{7}$$

- **RoI CLIP Loss**: Focuses on lesion regions for precise alignment between lesion image features and text features. Defined as:

$$\mathcal{L}_{\text{CLIP}_{roi}} = \frac{\mathcal{L}_{\text{CLIP}_{roi-cc}} + \mathcal{L}_{\text{CLIP}_{roi-mlo}}}{2} \tag{8}$$

- **Cross-View Consistency Loss**: Captures semantic consistency between CC and MLO views of mammograms. Defined as:

$$\mathcal{L}_{\text{cross}} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \cos\left(\hat{F}_{CC,i}, \hat{F}_{MLO,i}\right)\right) + \lambda\|\hat{F}_{CC,i} - \hat{F}_{MLO,i}\|^2 \tag{9}$$

The total loss combines these losses with weight coefficients:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{CLIP}_W} + \beta\mathcal{L}_{\text{CLIP}_{roi}} + \gamma\mathcal{L}_{\text{cross}} \tag{10}$$

Here, $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.2$ are the weight coefficients.

4 Experiments

We pre-trained the MM-CLIP model on image-region-text pairs using the Vindr-based data pipeline. Subsequently, we fine-tuned the visual encoder module of the pre-trained MM-CLIP model on datasets such as RSNA and MRDR using a classification data pipeline for the task of benign and malignant breast lesion classification. Compared to the original CLIP, our method demonstrates significant performance improvement.

4.1 Datasets

Vindr DataSet:A fully digital mammography dataset from Vietnam, significantly surpassing the scale of most existing publicly available datasets. Vindr comprises 20,000 images from 5,000 patients,

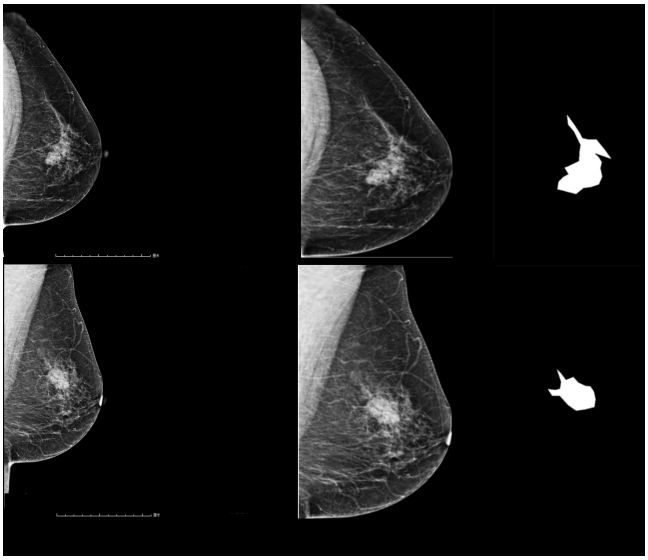

Figure 6: Data preprocessing

providing not only breast-level diagnostic information but also detailed region-level annotations. Vindr has a clear advantage over CBIS-DDSM in terms of the richness of lesion annotations and data scale.

RSNA Dataset:The RSNA-Mammo dataset, released by the Radiological Society of North America (RSNA) in 2023, is a mammographic imaging database aimed at advancing AI research in breast cancer detection. This dataset contains a large number of mammographic images, with over 55,000 images, covering a variety of breast abnormalities, including masses, calcifications, asymmetries, and architectural distortions.

MRMD: A private mammogram dataset, comprising medical reports, images, and RoI segmentation annotations, includes 1,150 images from 350 patients.

4.2 Data Preprocessing

During the pre-training phase, this study utilizes both the private MRMD dataset and the public Vindr dataset for model training. The RoI masks for the Vindr dataset are constructed based on annotation information, while the MRMD dataset directly provides images and corresponding masks. The original images are cropped to remove background information, resulting in a size of 480×760

In the downstream task testing phase, the RSNA dataset is introduced to validate the performance of MM-CLIP on long-tailed classification tasks. Experiments are conducted using an NVIDIA GeForce RTX 3090 GPU.During the pre-training phase, the visual encoder employs ResNet50, EfficientNet-B2, and EfficientNet-B4, initialized with ImageNet weights. The MRMD and Vindr datasets are trained for 60 and 30 epochs, respectively, with a batch size of 16. The AdamW optimizer is used (learning rate: 1e-6, weight decay: 1e-5, momentum: 0.9), and warm-up is applied to improve convergence.

4.3 Comparative Experiments

We compared the performance of various methods in benign and malignant classification on the MRMD dataset, with evaluation metrics including Accuracy and Area Under the Curve (AUC). The experimental results in Table 1 demonstrate that, compared to models pre-trained on ImageNet and several open-source methods, our approach can more effectively extract discriminative features, thereby significantly improving classification performance. When using EFB4 as the backbone of the visual encoder, our model achieves an improvement of 1.6% in Accuracy and 3.4% in AUC compared to the network pre-trained on ImageNet. These results fully validate the effectiveness of enhancing semantic associations through knowledge guidance.

Table 1: Comparison on the MRMD dataset.

Methods	Accuracy	AUC
ViT-S-16 (pretrained)	0.731	0.792
ResNet50 (pretrained)	0.756	0.813
ConvNet (pretrained)	0.761	0.810
EfficientNet-B4 (pretrained)	0.772	0.827
DenseNet[17]	0.780	0.844
Petrini et al. [7]	0.786	0.852
MM-CLIP+EFB4(ours)	**0.788**	**0.861**

The Vindr dataset is a highly imbalanced long-tailed dataset with a benign-to-malignant sample ratio of 49:1. This study uses this dataset to validate the model's ability to learn discriminative features and evaluate its classification performance under severe data imbalance. We selected two classic image-text multimodal models as baseline comparisons: CLIP (a vision-text contrastive learning multimodal model proposed by OpenAI) and ConVIRT [20] (a vision-text contrastive learning model proposed by Zhang et al. for the medical imaging domain). Since the original training data for these models are unavailable, we retrained them on the Vindr and MRMD datasets using their open-source training code to ensure the fairness of the experiments.

Table 2: Comparison on the RSNA dataset

Methods	Back bone	Zeroshot		Finetuning	
		Acc	AUC	Acc	AUC
CLIP	RN50	0.4813	0.4246	0.6375	0.7033
CLIP	EFb2	0.5700	0.5215	0.6757	0.7821
CLIP	EFb4	0.5936	0.5410	0.6931	0.7932
ConVIRT	RN50	0.5113	0.5246	0.6775	0.7233
ConVIRT	EFb2	0.5320	0.6015	0.7157	0.8011
ConVIRT	EFb4	0.5723	0.6310	0.7331	0.8349
MM-CLIP (Ours)	RN50	0.5734	0.6231	0.7234	0.8299
MM-CLIP (Ours)	EFb2	0.6188	0.6423	0.7620	0.8492
MM-CLIP (Ours)	EFb4	**0.6232**	**0.6537**	**0.7850**	**0.8593**

4.4 Ablation Study

We evaluated four different models on the RSNA validation set. Ablation experiments showed that the performance of the method without RoI CLIP loss significantly decreased (AUC: 0.7213 vs. 0.8563, Accuracy: 0.7213 vs. 0.784), demonstrating that RoI CLIP loss is a key module for improving performance. Secondly, the performance of the method without CLIP loss also significantly declined, validating its importance. The model without the cross-view matching module performed slightly worse than the full model. The baseline method showed a significant performance gap compared to the full model. Overall, RoI CLIP loss played the most significant role, while CLIP loss and the cross-view matching module further enhanced model performance. To further validate the impact of regional text knowledge features on model performance, we visualized the global similarity matrix and the regional similarity matrix. As shown in the figure7, the diagonal distribution is highlighted, representing matched positive samples, while low brightness indicates poor matching. It can be observed that the highlighted areas in the regional image-text similarity matrix are significantly fewer than those in the global matrix, indicating more precise image alignment.

Table 3: ablation study

CLIP	RoI CLIP	cross loss	**Accuracy**	**AUC**
✗	✗	✗	0.6867	0.7742
✓	✗	✗	0.7213	0.8153
✓	✓	✗	0.784	0.8563
✓	✓	✓	**0.785**	**0.8593**

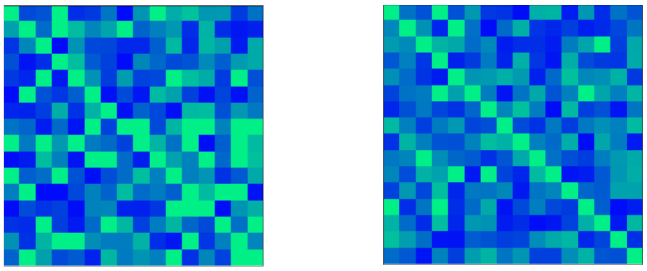

Figure 7: Similarity Matrix

5 Conslusion

In this work, we propose a novel pre-training framework for Mammography: Multi-granularity Mammography CLIP (MM-CLIP),to effectively address the semantic gap between image features and medical reports in mammography image-report pairs. A series of experiments conducted on the constructed mammography image pair dataset and mainstream public datasets fully validate the effectiveness of MM-CLIP in enhancing semantic associations, while also alleviating diagnostic challenges under data imbalance conditions. These experimental results demonstrate the potential of MM-CLIP in the field of mammography image analysis.

References

[1] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15016–15027.

[2] Yuanhong Chen, Hu Wang, Chong Wang, Yu Tian, Fengbei Liu, Yuyuan Liu, Michael Elliott, Davis J. McCarthy, Helen Frazer, and Gustavo Carneiro. 2022. Multi-View Local Co-Occurrence and Global Consistency Learning Improve Mammogram Classification Generalisation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 3–13.

[3] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?. In *Findings of the Association for Computational Linguistics: EACL 2023*. 1181–1193.

[4] Shantanu Ghosh, Clare B. Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. 2024. Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 632–642.

[5] Bowen Han, Luhao Sun, Chao Li, Zhiyong Yu, Wenzong Jiang, Weifeng Liu, Dapeng Tao, and Baodi Liu. 2024. Deep Location Soft-Embedding-Based Network with Regional Scoring for Mammogram Classification. *IEEE Transactions on Medical Imaging* (2024).

[6] Kshitiz Jain, Aditya Bansal, Krithika Rangarajan, and Chetan Arora. 2024. MM-BCD: Multimodal Breast Cancer Detection from Mammograms with Clinical History. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Vol. LNCS 15001. Springer Nature Switzerland, 144–154.

[7] D.G.P. Petrini et al. 2022. Breast Cancer Diagnosis in Two-View Mammography Using End-to-End Trained EfficientNet-Based Convolutional Network. *IEEE Access* 10 (2022), 77723–77731.

[8] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[9] L. Shen et al. 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports* (2019).

[10] X. Shu. 2020. Deep Neural Networks With Region-Based Pooling Structures for Mammographic Image Classification. *IEEE Transactions on Medical Imaging* (2020), 2246–2255.

[11] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. 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 (2021), 209–249. doi:10.3322/caac.21660

[12] A. Taha et al. 2022. Deep Is a Luxury We Don't Have. In *MICCAI*.

[13] Hung Q. Vo, Lin Wang, Kelvin K. Wong, Chika F. Ezeana, Xiaohui Yu, Wei Yang, Jenny Chang, Hien V. Nguyen, and Stephen T.C. Wong. 2024. Frozen Large-Scale Pretrained Vision-Language Models are the Effective Foundational Backbone for Multimodal Breast Cancer Prediction. *IEEE Journal of Biomedical and Health Informatics* (2024).

[14] Z. Wang et al. 2023. Dual-View Correlation Hybrid Attention Network for Robust Holistic Mammogram Classification. In *IJCAI*.

[15] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *arXiv preprint arXiv:2210.10163* (2022).

[16] T. Wei et al. 2022. Beyond Fine-Tuning: Classifying High Resolution Mammograms Using Function-Preserving Transformations. *Medical Image Analysis* (2022), 102618.

[17] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. 2019. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* (2019).

[18] Kihyun You, Suho Lee, Kyuhee Jo, Eunkyung Park, Thijs Kooi, and Hyeonseob Nam. 2022. Intra-Class Contrastive Learning Improves Computer Aided Diagnosis of Breast Cancer in Mammography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 55–64.

[19] Xiaofei Zhang, Yi Zhang, Erik Y. Han, Nathan Jacobs, Qiong Han, Xiaoqin Wang, and Jinze Liu. 2018. Classification of Whole Mammogram and Tomosynthesis Images Using Deep Convolutional Neural Networks. *IEEE Transactions on Nanobioscience* 17, 3 (2018), 237–242.

[20] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2022. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Machine Learning for Healthcare Conference*. PMLR, 2–25.

[21] W. Zhu, Q. Lou, Y.S. Vang, and X. Xie. 2017. Deep Multi-Instance Networks with Sparse Label Assignment for Whole Mammogram Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 603–611.