Information Fusion 118 (2025) 102998

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

MammoVLM: A generative large vision–language model for mammography-related diagnostic assistance

Zhenjie Cao [a,b], Zhuo Deng [a], Jie Ma [d], Jintao Hu [c,d], Lan Ma [a,*]

[a] Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, PR China
[b] AI Lab, Pingan Tech, Shenzhen, PR China
[c] Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Hong Kong
[d] Radiology Department, Shenzhen People's Hospital, Shenzhen, PR China

ARTICLE INFO

Keywords:
Mammogram
Multimodal foundation model
Vision–language model
Breast cancer
Medical Q&A
Diagnostic assistance

ABSTRACT

Inspired by the recent success of large language models (LLMs) in the general domain, many large multimodal models, such as vision–language models, have been developed to tackle problems across modalities.

In the realm of breast cancer, which is now the most deadly cancer worldwide, mammography serves as the primary screening approach for early detection. There is a practical need for patients to have a diagnostic assistant for their follow-up Q&A regarding their mammography screening. We believe large vision–language models have great potential to address this need. However, applying off-the-shelf large models directly in medical scenarios normally provides unsatisfactory results.

In this work, we present MammoVLM, a large vision–language model to assist patients with problems related to mammograms. MammoVLM has a sparse visual-MoE module that attends to different encoders based on the densities of the input image. Besides, we build a novel projection module, UMiCon, that leverages unimodal and multimodal contrastive learning training strategies to improve the alignment between visual and textual features. GLM-4 9B, an open-source LLM, is attached after previous multimodal modules to generate answers after supervised fine-tuning. We build our own dataset with 33,630 mammogram studies with diagnostic reports from 30,495 patients. MammoVLM has shown extraordinary potential in multi-round interactive dialogues. Our experimental results show that it has not only beaten other leading VLMs but also shows a professional capability similar to that of a junior radiologist.

1. Introduction

According to WHO, breast cancer has surpassed lung cancer as the world's number one cancer in terms of morbidity and mortality in 2020 [1]. Mammography screening is the most cost-effective method for early detection of breast cancer, with approximately 48 million mammograms performed annually in the U.S. It has been reported that the U.S. radiologists ranged from 66.7% to 98.6% for sensitivity and from 71.2% to 96.9% for specificity in mammogram-based breast cancer diagnosis [2]. During a patient's visit to the radiology department of a hospital, she usually has both mammography screening performed and a corresponding diagnostic report from the radiologist presented. Though this would normally be considered a closed cycle of one mammography diagnosis, patients still have concerns and uncertainties over their examination. This is why a diagnostic assistant is needed for follow-up Q&A. Patients may need a Chatbot with questions related to text only (language model) or related to mammograms (vision–language model).

Large language models (LLMs), such as ChatGPT, have recently achieved remarkable success, significantly enhancing applications like chatbots and AI agents. Their power stems from the ability to process and generate human-like text by analyzing vast amounts of data. These models are trained on a large corpus of text, which allows them to learn the statistical patterns and structures of language, enabling them to generate coherent and contextually appropriate text. Their strength lies in their ability to understand and produce human-like text [3,4].

On the other hand, multimodal large models, especially vision–language models (VLMs), have been developed to tackle problems beyond texts. They integrate computer vision and natural language processing (NLP) capabilities into a unified framework. These models can understand and generate descriptions of visual content and use natural language to guide the interpretation of visual data. This allows them to perform tasks that require an understanding of both visual and linguistic content, such as image captioning, visual question answering, and text-to-image generation. Vision–language models can

* Corresponding author.
E-mail address: malan@mails.tsinghua.edu.cn (L. Ma).

https://doi.org/10.1016/j.inffus.2025.102998
Received 3 December 2024; Received in revised form 15 January 2025; Accepted 29 January 2025
Available online 10 February 2025

transfer knowledge between modalities by learning the correlations between visual and linguistic data. VLMs can be applied to a wide range of practical tasks, such as helping visually impaired individuals understand the content of images, enhancing search engines with image and text queries, creating realistic images from textual descriptions, and more. The power of large vision–language models lies in their ability to bridge the gap between visual and linguistic information, enabling machines to understand and generate content in a way that is more similar to human cognition [5,6].

In this work, we aim to fully explore the cross-modality power of a large vision–language model, MammoVLM, to operate as a diagnostic assistant for Mammogram-related Q&A. Patient with her mammograms in hand often remains unsettled questions unsolved, for instance, whether and when should she follow up her screening [7]. With MammoVLM, patients can ask the questions related to their mammograms.

Since we do not find a similar mammography-related Q&A dataset, we collect our own dataset with 33,630 mammogram studies with diagnostic reports from 30,495 patients. All experiments and training are conducted within this dataset. In order to compare the performance of our model, we develop two sets of comparisons: open-source vision–language models and State-Of-The-Art LLMs connected after our visual-MoE + projection module (UMiCon). We design subjective and objective experiments for comparison. For the subjective part, thirty-eight patients are invited to provide their questions regarding their mammograms. We ask eight senior radiologists to evaluate answers generated by each model from aspects of correctness, rationality, helpfulness, and professionalism. As for the objective comparisons, we conduct ablation studies on different components of MammoVLM. Besides, to further prove the value of UMiCon as a projection module, we evaluate its classification performance separately with the designed pre-training task.

In conclusion, MammoVLM has four major contributions:

1. A sparse visual Mixture-of-Experts with three visual encoders, CLIP [8], ConvNeXt-Tiny [9], and Dinov2 [10] that processes mammograms with various densities. A tiny and efficient classifier will pre-process the mammograms and determine the correct encoder. Then, a visual representation will be transferred to the LLM before a projection module with better alignment.
2. A novel projection module, UMiCon, trained with unimodal and multimodal contrastive learning that bridges the gap between visual and textual modalities. In order to perform an ideal pre-training for UMiCon, we designed a classification task between BI-RADS 3 and 4, the most difficult BI-RADS to classify, to pre-train UMiCon. Due to the lack of high-quality Q&A pairs, we turn to mammograms and their existing corresponding diagnostic reports in this task.
3. We use GLM-4 9B [11] as the initial LLM. It is connected with the previous multimodal blocks and trained with supervised fine-tuning to generate final answers. Without training, it provides three answers with regard to the same question. Junior radiologists correct the common sense errors inside those questions, and senior radiologists pick the best answer based on its correctness and professional level. These Q&A pairs are the training data for the whole MammoVLM.
4. MammoVLM has not only beaten other leading VLMs by a noticeable margin but also shows a professional capability similar to that of a junior radiologist.

2. Related work

2.1. Large foundation models

Large foundation Models, encompassing architectures like Diffusion models [12], Transformer [13], BERT [14], DALL-E [15], GPT [16], GLM [11], and others [17–19], are grounded in deep learning and leverage transfer learning techniques. Often pre-trained or self-supervised, these models are generally trained on vast and varied datasets, enabling their versatility for a broad spectrum of subsequent applications through transfer learning. Their efficacy is largely due to the extensive scale of data they are trained on. In the realm of computer vision, large foundation models have advanced numerous image-related tasks, including reconstruction of natural images [20], object detection [18], segmentation [21], and classification [22]. They also hold significant potential in the field of medical imaging, where they can be applied to various imaging modalities such as X-rays [23], MRI, CT scans [19], and more. By training large foundation models on a wide range of medical data from different sources and modalities, these models can accumulate a wealth of medical knowledge, making them highly adaptable for multiple tasks within medical imaging.

2.2. Vision–language models

Large vision–language models (VLM) are part of large foundation models and have been a topic of intense research and development in recent years, with several notable works advancing the field [8,15,24–27]. CLIP [8] learns to match images to their corresponding text captions without using any classification labels and can be fine-tuned for a wide range of visual classification tasks simply by providing it with text prompts. DALL-E [15] generates images from textual descriptions. It combines a language model with an image generation model, allowing it to create images corresponding to complex text prompts. ALIGN [26] is pre-trained on a billion image-text pairs. It uses a simple and efficient approach for contrastive learning, which allows it to scale to such a large dataset. ALIGN performs strongly on various downstream tasks, including image classification, text-to-image retrieval, and zero-shot image classification. FLAVA [27] combines three pre-training objectives: language modeling, visual recognition, and joint image-text modeling. It aims to align and integrate representations from both modalities, achieving strong performance on various vision–language tasks. Besides, during the past year, VLMs based on pre-training research such as Qwen-VL [28], Yi-VL [29], CogCoM [30], CogVLM [31], BLIP-2 [32], LLaVA [33], and Vary [34] have been consistently pushing the performance boundaries across a spectrum of downstream tasks. These models and methods represent the cutting edge of vision–language research and are driving the development of applications. The crucial common focus of these methods is the alignment approach bridging the representation gap between modalities. Thus, one major contribution of our proposed MammoVLM is the novel projection module, UMiCon.

2.3. Large foundation models for medical research

Large foundation models for medical research are always trained/fine-tuned with clinical data, including digital medical imaging (X-rays, CT scans, MRI, etc.), user-upload images, clinical metadata, and clinical reports. Among them, vision–language models (VLMs) for VQA tasks account for a major part [35,36]. Image Content VQA (generating questions based on entity and table information in the medical reports) and Clinical Reasoning VQA (generating multiple-choice and short-answer questions related to disease diagnosis, staging, and treatment advice) are two mainstream tasks for VQA tasks. PubMedCLIP [37] is a dual-stream model based on CLIP, pre-trained on ROCO and fine-tuned on SLAKE and VQA-RAD for VQA. Similarly, BiomedCLIP [38] is based on CLIP, pre-trained on PMC-15, and fine-tuned on VQA-RAD and SLAKE for VQA. Med-Alpaca [39], a biomedical foundational model using DePlot or Med-GIT trained for visual interpretation and LLaMA-7B for language understanding, fine-tuned on medical question–answer pairs. HuaTuo tunes the LlaMA specifically with Chinese medical knowledge [40] . LLaVa-Med [41], on the other hand, is a single-stream

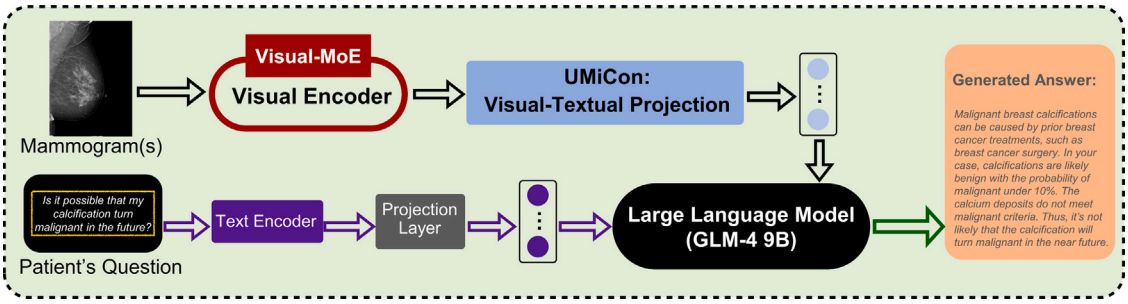

Fig. 1. Architecture of our MammoVLM. It consists of visual-MoE as the image encoder, UMiCon as the visual–textual projection module, and an LLM that generates the final answer.

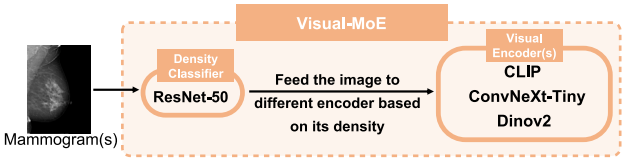

Fig. 2. Illustration of our Visual-Mixture-of-Experts (Visual-MoE). Based on the densities of input mammograms, ResNet-50 operates as a tiny selector of the encoders afterward.

model using CLIP for visual features and LLaMa-7B for language understanding, fine-tuned on PMC-15 for VQA. We follow the same fashion as these previous works, leveraging both mammograms and their corresponding diagnostic reports, and build MammoVLM to tackle mammography-related diagnostic questions.

2.4. Mammography screening

Previous works of mammography screening are mostly developed with mammograms only [42–47]. In contrast, we build a multimodal model with inputs from both mammograms and diagnostic reports, namely UMiCon. UMiCon operates as the projection module within MammoVLM and is pre-trained with the classification task between BI-RADS 3 and 4. To best leverage information from both modalities, we first set up an unsupervised learning stage to improve the model's representation ability as in Fig. 3. During this stage, we conduct contrastive learning for both unimodal and multimodal branches, whose effectiveness has been widely proved in recent studies on medical images [48–54]. Then, the supervised learning stage carries on. In this stage, we design a simple yet effective cross-modal fusion strategy. Also, we add a cross-modal contrastive learning head to search for an ideal embedding space. It further improves the UMiCon's learning ability over these fused features.

3. Methodology

Overall, the MammoVLM is a VQA system that provides answers based on questions and mammograms from the patients. Fig. 1 demonstrates the overall framework of MammoVLM. It has a sparse visual-MoE that processes the mammograms and generates visual features. A projection module UMiCon, pre-trained in a specific multimodal task, aligns visual features with textual features. GLM-4 9B [11] generates the final answers. Note that the same LLM without any fine-tuning is also used to prepare training and testing data in the first place, which will be explained in Section 4.1.1.

3.1. Visual-MoE

There are different density categories for mammograms: being almost entirely fatty (A), having scattered areas of dense fibroglandular breast tissue (B), having many areas of glandular and connective tissue (C), or being extremely dense (D) [55]. Due to the variation among these densities, it is practical to select different encoders for deep learning models based on different densities [56]. The sparse Mixture-of-Experts (MoE) architecture offers a scalable, efficient, and flexible solution for encoder selection, enabling the construction of large foundation models while preserving computational efficiency. Thus, we adopt a visual-MoE to process the mammograms, as illustrated in Fig. 2. We experiment with different visual encoders and select three which are most sensitive to densities. Based on Table 6, CLIP [8] performs best for category A breast, ConvNeXt-Tiny [9] generates the best features for category C and D, and Dinov2 [10] is best for category B. Thus, we add a tiny selector, ResNet-50, functioning as a density classifier to process the input mammograms first before sending them to one of the visual experts. We train ResNet-50 with the labels of densities within our dataset. After determining the density category of the input mammogram, it proceeds the image into one of the encoders. The results of the sparse MoE as well as the accuracy of classifier ResNet 50 have been presented in the Table 6 and Section 5.2.

3.2. Pre-training the projection module UMiCon

The visual features generated by Visual-MoE should be aligned with textual features in order to be sent into LLM afterward. Developing a projection module to complete this alignment is of great significance. In this section, we explain how we utilize both **U**nimodal and **M**ulti**m**odal **Con**trastive learning to pre-train a projection module, UMiCon.

3.2.1. Pre-training task selection: BI-RADS 3 vs. 4

Collecting enough amount of Question–Answer pairs related to mammography screening for pre-training is highly unpractical. The consistency of Q&A pairs is also hard to maintain. As we have collected a sufficient amount of mammograms with their corresponding diagnostic reports, which are also related to textual data, we leverage **mammogram–report pairs** to pre-train UMiCon as our projection module.

When designing the pre-training task for UMiCon, we target the most challenging BI-RADS classification task, BI-RADS 3 vs. 4. According to the ACR's Breast Imaging Reporting and Data System (BI-RADS) criteria [57], breast lesions in mammography are divided into BI-RADS 0~6; BI-RADS 0 is an incomplete assessment; BI-RADS 1 finds no lesions during the examination, and its malignancy probability is almost 0, which is the same as BI-RADS 2; BI-RADS 5 lesions have malignancy possibilities of ≥ 95%; lesions confirmed to be malignant by biopsy are classified as BI-RADS 6. Unlike these BI-RADS categories, the benignity or malignancy is almost certain; BI-RADS 3 and 4 bring many more diagnostic difficulties and inconsistencies between diagnosers. BI-RADS 3 refers to a malignancy possibility of ≤ 2%; BI-RADS 4 lesions have a malignancy probability of 2% to 95% and are divided into three subtypes: 4 A, 4B, and 4C. Nearly all misclassifications for mammography screening happen between BI-RADS 3 and 4. In real

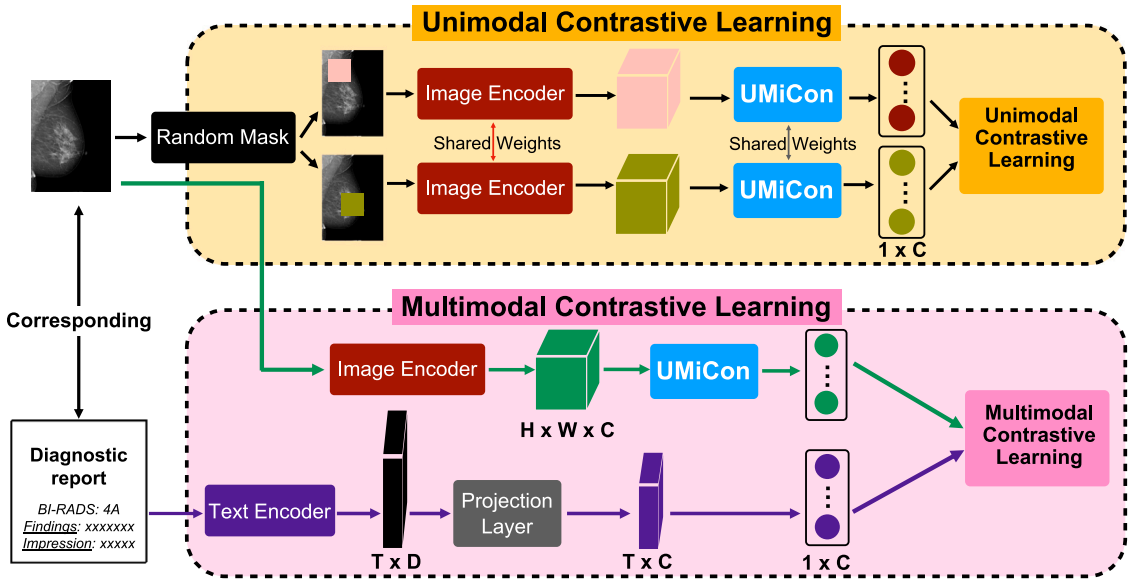

Fig. 3. UMiCon pre-training: *Unsupervised learning* stage.

clinical scenarios, 0.9% to 7.9% of suspected benign lesions evaluated in BI-RADS 3 are diagnosed as malignant after surgery [58,59]. On the other hand, 70% of the false positives in mammography happen from BI-RADS 4. This brings problems of over-diagnosis and over-treatment.

The challenge of distinguishing between BI-RADS 3 and 4 underscores the value of this classification task. To address this, we formulate the task in a multimodal way, leveraging both mammograms and diagnostic reports from these two BI-RADS categories. This method enhances the capability of our projection module to bridge the gap between visual and textual data modalities. Consequently, we focus on pre-training UMiCon specifically for the classification of BI-RADS 3 vs. BI-RADS 4. The motivation and importance of this task are discussed in Section 1 and 4.1.3. We dedicate Table 7, Table 8, and Section 5.3 to present results of UMiCon on this pre-training task and its effectiveness.

3.2.2. UMiCon structure

UMiCon has a pooling layer and flatten operation that maps the visual features to a six-layer MLP within it. The overall framework of the UMiCon's pre-training is illustrated in the below figures. Fig. 3 shows the *unsupervised learning stage*. The resulting image encoder, text encoder, projection layer, and UMiCon are then transferred to the *supervised learning stage* as presented in Fig. 4. We further elaborate on both phases in the following subsections.

3.2.3. Unsupervised learning stage

As shown in Fig. 3, during this stage, inputs are mammograms and their corresponding diagnostic reports pairs from the same patients. Two branches are carried out simultaneously, one responsible for unimodal contrastive learning and the other for multimodal contrastive learning.

Unimodal contrastive learning branch The input mammogram is sent into the *random mask* module, in which a mask accounting for 8% of the image area is randomly added twice, generating a pair of input images. The mask is a square and locates strictly inside the input image. Then, the branch is carried out by the Siamese contrastive learning module, which consists of a Siamese encoder (red block in Fig. 3) and a Siamese projection module (UMiCon in Fig. 3). The pair of input mammograms are simultaneously fed into the shared-weight encoders. UMiCon then projects the encoded features into two 1-D vectors representing their class's likelihood. The contrastive loss [60] is designed to draw the samples from the same class closer and separate the samples from different classes farther apart in the projected space. Given a pair of input images (I, I'), we use the regular L2 distance in the loss function and set *margin* as 1:

$$L(I, I') = \begin{cases} D^2 & \text{if } l_I = l_{I'} \\ \max(0, margin - D)^2 & \text{if } l_I \neq l_{I'}, \end{cases} \tag{1}$$

where

$$D = \left\| \mathcal{P}_{\text{sia}}(\mathcal{E}_{\text{sia}}(I)) - \mathcal{P}_{\text{sia}}(\mathcal{E}_{\text{sia}}(I')) \right\|_{L2}, \tag{2}$$

and $\mathcal{E}_{\text{sia}}(\cdot)$ and $\mathcal{P}_{\text{sia}}(\cdot)$ denote the Siamese encoder and Siamese projection module, respectively; l_I and $l_{I'}$ indicate the corresponding BI-RADS labels. The loss for a batch of N image pairs can be simply defined as $\mathcal{L}_{\text{batch}} = \sum_{i=1}^{N} L(I_i, I_i')$. In this case, *samples always belong to the same class* as they are augmented from the same image. The image encoder here is ConvNeXt-Tiny [9].

Multimodal contrastive learning branch The same input image, as that of the unimodal branch, will pass through the image encoder and UMiCon sequentially, which share weights with those in the unimodal branch. The corresponding diagnostic report is encoded by a text encoder MacBERT-base [61], a satisfying text encoder specifically for Chinese. The extracted text feature is then linearly projected by the projection layer (gray block in Fig. 3) to dimension C, consistent with the image feature. These two 1-D vectors represent different modalities, images and text. Lastly, multimodal contrastive learning across these vectors operates with the same algorithm described in the unimodal branch.

3.2.4. Supervised stage

The inputs of this stage are mammogram-diagnostic report pairs from various patients in our dataset, of which BI-RADS labels are available. The image encoder, text encoder, projection layer, and UMiCon (red, purple, gray, and black blocks in Fig. 4) are transferred from the unsupervised learning stage.

Cross-modality fusion Image encoder and text encoder with UMiCon and projection layer set features from the mammogram and diagnostic report to the same dimension. After the straightforward concatenation and layer normalization, the cross-modality feature is fused into a 2D vector. This is a modality fusion pipeline with no attention-based calculation, though previous works have utilized the transformer and multi-head cross-attention [52,53]. Experimental results in Table 8 show that this simple yet effective approach is good for generating promising results.

Classification head and cross-modal contrastive learning The cross-modality 2-D vector is trained through a classification head,

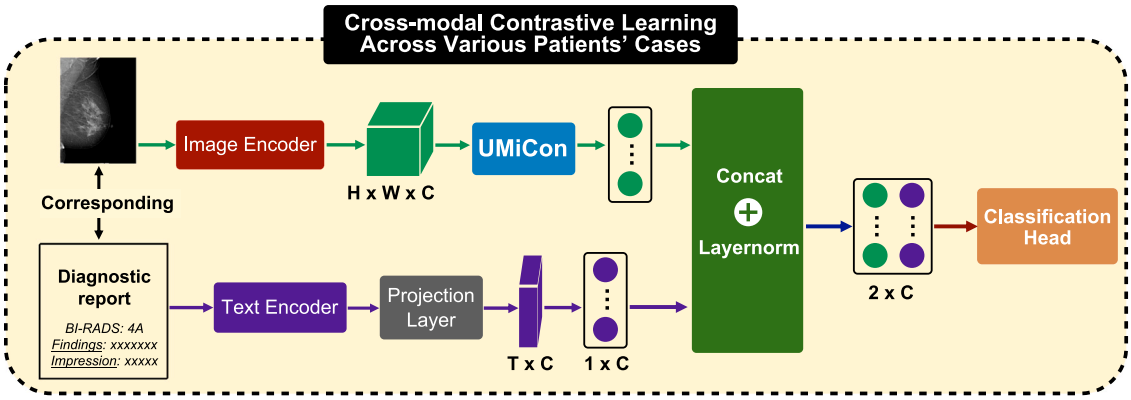

Fig. 4. UMiCon pre-training: *Supervised learning* stage.

which contains Sigmoid and cross-entropy loss calculated with the BI-RADS labels of each patient. Besides, given the subtle and hard-to-distinguish differences between BI-RADS 3 and 4, we add another cross-modal contrastive learning head upon the cross-modality 2-D vector. A pair of patients' data, in this case, a pair of 2-D vectors, are being calculated with each other using contrastive loss. From our experiments in Table 8, this contrastive learning head boosts the model's classification performance by improving its representation ability with multimodal BI-RADS 3 and 4 data.

3.3. LLM tuning

We employ the Supervised Fine-Tuning (SFT) strategy to fine-tune our LLM. During the fine-tuning process, we freeze the visual encoder and UMiCon module parameters. We choose GLM-4 9B [11] as the initial LLM, which is well within the capacity of our GPU resources. In the SFT stage, the initial inputs to our MammoVLM (mammograms and patient questions) should be converted into an instruction-aware format consisting of instructions, visual input information, and ground-truth responses. The details of training samples are discussed in Section 4.1.1. Through instruction tuning, the LLM can predict the answer given the instruction and visual input information:

$$A = f(I; M; \theta) \tag{3}$$

where A represents the answer output by LLM, θ denotes the LLM's parameters, $I \in \mathbb{R}^{1\times1\times128}$ denotes the word embedding vector output by the projection layer, and $M \in \mathbb{R}^{1\times1\times128}$ is the visual feature vector mapped from the visual space to the textual space by the UMiCon module.

The training objective is typically the original auto-regressive objective used to train LLMs [62], based on which the LLM is encouraged to predict the next token of the response. The objective can be expressed as:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} log\ p(R_i|I, R_{<i}; \theta) \tag{4}$$

where R represents the ground-truth response, and N is the length of the ground-truth response. We have also experimented with other LLMs and results have been shown in Section 5.1 and Table 2.

4. Experiment design

4.1. Data collection

Our data is collected from three collaborative hospitals at distinct geographical locations using Siemens and Giotto equipment following the ACR standard from 2011 to 2021.[1] A large-scale dataset containing 33,630 mammogram studies with diagnostic reports from 30,495 patients was collected. The distribution is as follows: BI-RADS 1 with 19,698 cases, BI-RADS 2 with 8,857 cases, BI-RADS 3 with 3,636 cases, BI-RADS 0 with 393 cases, BI-RADS 4 with 821 cases, BI-RADS 5 with 182 cases, and BI-RADS 6 with 43 cases. The benign and potentially malignant distribution based on these categories is 32,348 for benign and 1,282 for potentially malignant cases. Each breast's standard mammography screening case has two X-ray projection views: a craniocaudal (CC) view and a mediolateral oblique (MLO) view. The image-level labels are extracted from the diagnosis reports. All the reports we collected have already been reviewed by at least a senior radiologist. Notably, during the pre-training stage of UMiCon as in Section 3.2, we have blanked the BI-RADS information initially contained in the reports. All data has been pre-processed and does not contain any personal or sensitive information about the patient.

4.1.1. Supervised fine-tuning

We set radiologists into two groups, junior and senior, separated by their years of expertise. Junior imaging physicians have 5~6 years of experience, while senior physicians have 21 years of working experience and professional breast imaging training. We record 3,753 questions from relevant patients. We use GLM4-9B as the initial language model (ILM) to provide three answers to each question, where the mammograms serve as a prompt for ILM. Junior radiologists corrected the common sense errors inside those answers if there were any. Then, senior radiologists picked the best answer based on its correctness and professional level for each patient's case. Each training sample consisted of a question (from the patient), her mammogram(s), and an answer (picked by the senior radiologist). All these training samples are leveraged in the LLM tuning stage as in Section 3.3.

4.1.2. Visual-MoE training

As described in Section 3.1, ResNet-50 operates as a selector for the visual encoders afterward based on the densities of input mammograms. Thus, we train the ResNet-50 with density labels. Among the 33,630 mammogram studies we collected, categories A, B, C, and D have accounted for 1,680, 8,398, 11,786, and 11,768 studies, respectively. The density labels are acquired from their existing diagnostic reports.

[1] Ethics and institutional review board have approved this retrospective case-control study.

Table 1
Comparison between MammoVLM and Various SOTA VLMs (Average Scores on four aspects are from Eight Senior Radiologists). *Junior* and *Senior* represent radiologists with 5 and 21 years of working experience, respectively.

Aspects	Correctness	Rationality	Helpfulness	Professionalism	Completeness
LLaVA-1.5 [33]	81.3	77.8	91.0	80.6	96.6
CogVLM [31]	64.6	75.6	86.5	78.6	88.7
QWen-VL [28]	86.5	74.2	84.7	88.7	90.7
Mammo-CLIP [63]	80.4	81.2	88.9	89.4	79.4
BiomedGPT [38]	91.5	80.3	86.4	88.8	90.8
MammoVLM	**91.5**	**82.6**	**91.2**	**94.6**	**96.5**
Junior	90.6	83.0	93.4	91.1	93.0
Senior	92.7	86.1	96.3	95.1	97.1

Table 2
Performance comparison using different LLMs (Average Scores on four aspects are from Eight Senior Radiologists).

Method	Correctness	Rationality	Helpfulness	Professionalism	Completeness
Llama-3 [64]	87.5	78.7	89.0	89.6	94.9
Mixtral 8 × 22B [65]	84.6	79.6	90.6	88.6	94.1
QWen [28]	82.8	72.4	83.2	85.7	96.1
GLM-4 9B (Ours)	**91.5**	**82.6**	**91.2**	**94.6**	**96.5**

Table 3
Ablation studies on Visual-MoE, UMiCon, and LLM SFT of MammoVLM. (Average Scores on four aspects are from Eight Senior Radiologists).

Method	Correctness	Rationality	Helpfulness	Professionalism	Completeness
CLIP only w/o MoE + Linear Mapping + LLM (freeze)	71.5	67.3	69.0	65.8	80.8
ConvNeXt only w/o MoE + Linear Mapping + LLM (freeze)	70.8	70.3	71.4	68.9	86.9
Visual-MoE + Linear Mapping + LLM (freeze)	77.9	75.4	76.3	74.0	88.4
Visual-MoE + UMiCon + LLM (freeze)	86.4	78.8	88.1	90.2	92.6
Visual-MoE + UMiCon + LLM SFT **(MammoVLM)**	**91.5**	**82.6**	**91.2**	**94.6**	**96.5**

Table 4
Number of patients in each BI-RADS category for UMiCon training.

BI-RADS		3	4A	4B	4C	Total
Pathology	*Negative*	990	1195	281	18	2484
	Positive	189	462	652	770	2073
Total		1179	1657	933	788	4557

Table 5
Three training settings for various stages of MammoVLM.

Training settings		Batch size	GPU	Initial learning rate
Visual-MoE		6	4	2×10^{-4}
UMiCon	*Unsupervised*	12	6	2×10^{-5}
	Supervised	24	2	1×10^{-5}
LLM tuning		24	24	1×10^{-5}

4.1.3. Projection module UMiCon

The value of identifying BI-RADS 3 and 4 mammograms In clinical scenarios, nearly all misclassifications of mammography screening happen between BI-RADS 3 and 4. On the one hand, according to the BI-RADS guideline [57], the malignancy possibility of BI-RADS 3 is ≤ 2%. Therefore, radiologists define them as benign lesions. However, in actual clinical work, 0.9% to 7.9% of suspected benign lesions evaluated in BI-RADS 3 are diagnosed as malignant after surgery. On the other hand, 70% of false positives in mammography examinations are BI-RADS 4. These false alarms from BI-RADS 4 result in unnecessary diagnosis and over-treatment. The certain degree of false negatives and positives in BI-RADS 3 and 4 diagnoses further prove the clinical importance of this classification task.

Sub-dataset for UMiCon Training We create a sub-dataset to train UMiCon. This dataset contains each patient's images, diagnosis reports, and corresponding biopsy results. The training and validation sets include 4,557 patients, among which 1,179 are diagnosed as BI-RADS 3 and 3,378 as BI-RADS 4. Table 4 shows the number of patients in each BI-RADS category. Among the 1,179 initially evaluated as BI-RADS 3 cases, 189 (16.03%) were later proved malignant by biopsy results. As for BI-RADS 4, 1,914 of the 3,378 (56.66%) BI-RADS 4 patients were later proved benign by biopsy. According to Lee [66] and Orel et al. [67], biopsy-proven benign and malignant cases should be categorized as BI-RADS 3 and 4, respectively. Thus, we correct the BI-RADS labels, assigning 2,484 cases as BI-RADS 3 and 2,073 cases as BI-RADS 4. The dataset is split into the training and validation sets by 8:1 ratio. Our test set includes 601 patients collected within 61 consecutive days (Mar∼Apr 2019) from one of those three hospitals, among which 445 are BI-RADS 4 and 156 are BI-RADS 3 patients. All data in our dataset come with biopsy-proven results.

4.2. Implementation details

Visual-MoE, UMiCon, and LLM tuning share the following training parameters. All input images are resized to 1008 × 800 and retain the original aspect ratio. C is 128. 6 warming-up steps are added after the start. The initial learning rates are reduced by a factor of 10 after 100 epochs. Adam is used [68], with a weight decay of 5×10^{-4}. NVIDIA A100 GPUs (40G memory each) are used. The model training normally completes within 280 epochs. All implementations are with Python 3.11.0 and PyTorch 1.9.0. The detailed training settings of batch size, GPU usage, and initial learning rate in different stages of MammoVLM are illustrated in the Table 5. During inference, the model's runtime is less than 5 s with 4 A100 GPUs. We have compared several LLMs,

4.3. Evaluation metric

We compare the performance of MammoVLM with five open-source SOTA VLMs: LLaVA-1.5 [33], CogVLM [31], QWen-VL [28], BiomedGPT [38], and Mammo-CLIP [63]. We apply the same training data and SFT recipes to build an equal comparison. Also, we invite a junior (5 years of working experience) and senior radiologist (21 years of experience) to join the experiments. Besides, we apply three open-source SOTA LLMs, Llama-3 [64], QWen [28], Mixtral 8 × 22B [65], to replace our GLM-4 9B to prove the universality of our visual-MoE and

UMiCon design in a broad selection of vision–language models.

Details on the experiments for MammoVLM will be discussed in Section 5.1

4.3.1. Visual-MoE

We evaluate several mainstream visual encoders regarding their classification ability towards each density. The area under the receiver operating characteristic curve (AUC) is applied to select the three best encoders for each density degree, with which we evaluate the tiny selector of encoders, ResNet-50, too.

4.3.2. Projection module UMiCon

To fully prove the value of UMiCon as a projection module, we conduct objective comparisons in the task of BI-RADS 3 and 4 classifications, the value of which has been discussed in Section 4.1.3. We re-implement methods as in [44–46], the previously reported SOTA methods for mammography screening, and apply them to our datasets. We compare the system performance using quantitative metrics, including area under the receiver operating characteristic curve (AUC), accuracy (ACC), sensitivity (SEN), specificity (SPC), positive predictive value (PPV), and negative predictive value (NPV), where default thresholds are set as 0.5. The ground truth of this experiment is the biopsy results, where a benign BI-RADS 3 and a malignant BI-RADS 4 are true negative and true positive, respectively.

5. Experimental results and discussion

5.1. Performance evaluation of MammoVLM

5.1.1. Comparison with other SOTA VLMs and human doctors

We invite forty-eight patients to provide their questions regarding their mammograms. We request twelve senior radiologists to evaluate the answers generated by each VLM model based on correctness, rationality, helpfulness, professionalism, and completeness, which are the most crucial aspects of medical VQA, according to radiologists and researches [69,70]. Each of these metrics is scored on a scale from one to one hundred, with one being the worst and one hundred being the best. The results we have presented in Table 1 are the best ones, with other VLMs, BLIP2 [32] and Yi-VL [29], being omitted due to not being competitive enough.

Note that the same training data and SFT recipes are applied to these VLMs in order to build this fair comparison. From Table 1, we can see that MammoVLM outperforms all other competitors by a noticeable margin. Completeness shows the only exception as LLaVA [33] beats MammoVLM by a narrow margin. LLaVA has proved to be comprehensive on a variety of text generation tasks [33]; this result further proves it as a reliable model. As for correctness, BiomedGPT [71] joins MammoVLM to score the highest but falls far behind MammoVLM in the other three aspects. Considering that all these comparing VLMs are the current leading open-source models, we believe MammoVLM has superior and promising capability in this task of Mammography-related Q&A.

Additionally, we invite junior and senior radiologists to provide their answers to the same questions and mammograms as for these VLMs. With 21 years of working experience, the senior radiologist outperforms all competing VLMs, including MammoVLM. However, MammoVLM does show extremely competitive performance compared with the junior radiologist with 5 years of working experience, outperforming her in terms of correctness and professionalism. Thus, we believe that MammoVLM may already have the professional capabilities of a junior doctor in this mammography-related Q&A task.

Table 6
AUC comparison of different encoders in VisualMoE for classification performance on breast densities categories A, B, C, and D.

Method	A	B	C	D
ResNet-50 [72]	0.68	0.71	0.84	0.85
CLIP [8]	**0.88**	0.85	0.79	0.89
DINOv2 [10]	0.76	**0.90**	0.86	0.81
ConvNeXt-base [9]	0.81	0.79	0.82	0.85
ConvNeXt-tiny [9]	0.84	0.88	**0.93**	**0.95**

5.1.2. MammoVLM with other SOTA LLMs

We replace the GLM-4 9B in MammoVLM with other leading open-source LLMs and present results in Table 2. The same training data and SFT training settings are applied to them as well. Though a performance gap exists between these competing LLMs and GLM-4 9B, it is not significant enough to rule out these competing LLMs completely. On the other hand, this proves the universality of our visual-MoE and UMiCon design. We believe that adopting the overall design of the framework of MammoVLM may be successful regardless of the LLM selection for another medical VQA task.

5.1.3. Ablation study on MammoVLM

Table 3 is the ablation experiments of MammoVLM. The first two rows represent the single vision encoder instead of MoE, vanilla linear mapping instead of UMiCon, and GLM-4 9B without SFT. Row 3 adds Visual-MoE and improves in all aspects.

The most noticeable performance jump appears in row 4, where UMiCon is applied as a projection module between visual and textual features. This proves that the projection module is crucial to the success of vision–language models. We will further discuss UMiCon's effectiveness in Section 5.3.

Row 5 presents the best performance by adopting supervised fine-tuning for the LLM, which is also our SOTA MammoVLM. This SFT, with our collected training Q&A pairs, pushes this foundation model towards its optimal point.

5.2. Visual-MoE results

Table 6 illustrates the classification performance of different vision encoders on breast density categories A, B, C, and D. AUC results have been presented. CLIP [8], DINOv2 [10], and ConvNeXt-tiny [9] have each topped the classification ability on breast density categories A, B, and C &D. The results show that it is practical to use different encoders for different density mammograms, according to their classification performance. This motivates us to adopt the MoE architecture as in Section 3.1 instead of a single vision encoder. Using data in Section 4.1.2, ResNet-50 achieves an AUC of 0.94 to classify breast densities and operates as the selector of different encoders.

5.3. Effectiveness of umicon's pre-training strategy

As explained in Section 3.2, we design the overall framework of UMiCon's pre-training with a BI-RADS 3 vs. 4 classification task. In this section, we evaluate the effectiveness of our approach under this classification task. Table 7 shows the performance comparison with different approaches. Note that **UMiCon** in this section represents our pre-training strategy of UMiCon as in Section 3.2. We invite four doctors to our experiments: junior imaging physicians with 5 and 6 years of experience, respectively, A and B, and two senior physicians, C and D, with 21 years of working experience and professional breast imaging training. Collaborative doctors conduct the experiments solely and with the help of UMiCon.

Comparison with the State-of-the-Arts As shown in Table 7, compared with previous SOTA methods [44–46], the UMiCon framework effectively outperforms them in all metrics, including the AUC and

Table 7
Performance comparison of different approaches to identify BI-RADS 3 and 4, including previously reported best methods in the top 3 rows. *A, B, C,* and *D* are doctors with different professional levels. The 95% confidence intervals (CI) are shown in the square brackets.

Method	AUC	ACC	SEN	SPC	PPV	NPV
Yala et al. [44]	0.68 [0.62, 0.74]	0.67	0.72	0.59	0.61	0.72
Cao et al. [45]	0.61 [0.56, 0.66]	0.60	0.81	0.45	0.56	0.70
Mckinney et al. [46]	0.64 [0.62, 0.66]	0.64	0.78	0.49	0.60	0.77
UMiCon (ours)	**0.70** [0.65, 0.75]	**0.69**	**0.81**	**0.59**	**0.62**	**0.79**
A	0.62 [0.56, 0.69]	0.62	0.80	0.45	0.58	0.70
A + UMiCon	0.74 [0.68, 0.80]	0.72	0.91	0.58	0.61	0.89
B	0.61 [0.54, 0.68]	0.60	0.83	0.38	0.56	0.70
B + UMiCon	0.75 [0.69, 0.81]	0.73	0.86	0.64	0.64	0.86
C	0.69 [0.63, 0.76]	0.67	0.88	0.51	0.57	0.85
C + UMiCon	0.76 [0.7, 0.82]	0.76	0.92	0.61	0.69	0.89
D	0.71 [0.65, 0.78]	0.68	0.94	0.48	0.57	0.92
D + UMiCon	0.78 [0.72, 0.84]	0.77	0.90	0.65	0.71	0.88

Table 8
Ablation study for components in UMiCon. *Base-M* is trained without any contrastive learning. *Uni-M*, *Multi-M*, and *Cross-M* represent Unimodel, Multimodel, and Cross-model, respectively. **Row 4 is UMiCon.**

Method	AUC	ACC
Base-M (Supervised classification)	0.56	0.6
Base-M + Uni-M	0.61	0.63
Base-M + Uni-M + Multi-M	0.64	0.66
Base-M + Uni-Ml + Multi-M + Cross-M	**0.70**	**0.69**
Base-M (MHCA [53]) + Uni-M + Multi-M + Cross-M	0.69	0.67

Table 9
Detailed description of MammoVLM benign and malignant test data.

Mammograms	BI-RADS	Count	Total
Malignant Mammograms	4	88	102
	5	6	
	6	8	
Benign Mammograms	1	1346	2300
	2	728	
	3	226	
Excluded	0	136	136

Table 10
Zero-Shot Results of the mammographic malignancy screening task with different VLMs. The 95% confidence intervals (CI) are shown in the square brackets.

Method	AUC	Spe. (Sen. = 20%)
LLaVA-1.5 [33]	0.8671 [0.8667, 0.8671]	0.9405 [0.9401, 0.9411]
BiomedGPT [71]	0.8946 [0.8940, 0.8952]	0.9701 [0.9695, 0.9707]
Mammo-CLIP [63]	0.9071 [0.9069, 0.9074]	0.9786 [0.9779, 0.9795]
QWen-VL [28]	0.9070 [0.9065, 0.9075]	**0.9808 [0.9804, 0.9811]**
MammoVLM	**0.9270 [0.9264, 0.9276]**	**0.9902 [0.9899, 0.9905]**

accuracy, and represents a promising overall performance. Regarding sensitivity and specificity, UMiCon surpasses the previous SOTA methods by 4% and 8%, on average, claiming it can effectively identify BI-RADS 3 and 4 without sacrificing one another.

Comparison with the doctors The AUC (0.70) of standalone UMiCon is higher than that of junior doctors (A, B) and close to professional doctors (C, D). Every doctor's AUC performance improved over themselves when combined with the UMiCon to distinguish BI-RADS 3 and 4.

Ablation study on UMiCon To investigate the effectiveness of different modules and stages inside UMiCon, we further break them down into ablation experiments. Table 8 shows the ablation study results. The base model is trained with our cross-modal fusion without any contrastive learnings (row 1). After conducting unimodal contrastive pre-training with image inputs, the model's performance improves to par with junior doctor (row 2). Multimodal contrastive pre-training further increases both the AUC and accuracy by 3% (row 3). Cross-modal contrastive learning sees the most significant rise of AUC by 6% (row 4), which is the level of our SOTA UMiCon. In addition, in order to validate the effectiveness of our designed cross-modal fusion strategy as in Section 3, we replace it with multi-head cross attention and transformer architecture as in [53] (row 5). Although row 5's performance is slightly below UMiCon, UMiCon's cross-modal fusion design is much simpler and requires no attention computation.

5.4. Zero-shot results on identifying malignant mammograms

In order to validate the effectiveness of MammoVLM, we test its zero-shot results on mammographic malignancy screening tasks. Table 9 shows a new dataset that MammoVLM has not seen during training. We simplify this task by gathering BI-RADS 1∼3 into the non-malignant class and BI-RADS 4∼6 into the malignant class, according to [73,74]. BI-RADS 0 is excluded due to its uncertainty of malignancy. We inference MammoVLM by sending it the test image and questions about its malignancy. The performance comparison among various VLMs is shown in Table 10, where the AUC values are for the non-malignant class, and a sensitivity of 20% means that 20% non-malignant mammograms are confidently screened out from all non-malignant mammograms. Although all these VLMs show strong generalization capabilities, MammoVLM still outperforms others by adapting to new, unseen scenarios. By analyzing the zero-shot results, we believe MammoVLM has strong potential for novel tasks and further real-world applications.

6. Conclusion

Large foundation models have proven successful in a wide range of applications and have great potential in the medical research domain. This work presents MammoVLM, a vision–language large model that solves mammography-related diagnostic assistance problems. This VQA system consists of three major parts: a visual-MoE that encodes the input mammograms based on their densities with different visual experts, a projection module UMiCon that is pre-trained with unimodal and multimodal contrastive learning on a challenging classification task BI-RADS 3 vs. 4, and supervised fine-tuning on LLM that generates the final answers. Subjective experimental results on our large-scale dataset have shown that MammoVLM has not only beaten leading open-source VLMs but also shows competitive professional capabilities at a junior radiologist level. Besides, objective ablation results on MammoVLM further prove the effectiveness of different parts within MammoVLM. Separately, we elaborate on the experiments for our projection module UMiCon on the classification task BI-RADS 3 vs. 4. Given the projection module UMiCon's demonstrated strength in distinguishing between BI-RADS 3 and 4, we plan to explore its potential for enhancing BI-RADS classification across all categories through fine-tuning in our future work.

Mammography is the primary screening method for early detection of breast cancer, the world's most fatal and deadly cancer. Leveraging the large foundation model MammoVLM to serve as mammography-related diagnostic assistance has huge potential to improve medical service efficiencies and effectiveness.

CRediT authorship contribution statement

Zhenjie Cao: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Zhuo Deng:** Visualization, Validation, Project administration, Methodology, Formal analysis, Data curation. **Jie Ma:** Investigation, Funding acquisition. **Jintao Hu:** Investigation, Funding acquisition. **Lan Ma:** Visualization, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by Ping An Technology (Shenzhen) Co., Ltd., China and PAII Inc. This work is funded by Shenzhen Science and Technology Innovation Bureau, China under GJHZ20220913142613025. This work is also funded by Guangdong Medical Research Foundation, China under A2024506. Zhenjie Cao and Zhuo Deng contribute equally to the article.

Data availability

Data will be made available on request.

References

[1] World Health Organization, Breast cancer, 2020, URL http://www.who.int/news-room/fact-sheets/detail/breast-cancer.
[2] C.D. Lehman, R.F. Arao, B.L. Sprague, J.M. Lee, D.S. Buist, K. Kerlikowske, L.M. Henderson, T. Onega, A.N. Tosteson, G.H. Rauscher, et al., National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium, Radiology 283 (1) (2017) 49–58.
[3] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for large language models: A survey, ACM Trans. Intell. Syst. Technol. 15 (2) (2024) 1–38.
[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–45.
[5] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[6] W. Dai, J. Li, D. Li, A.M.H. Tiong, J. Zhao, W. Wang, B. Li, P.N. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2024).
[7] M.W. Kissin, A. Subramanian, To follow-up or not to follow-up, that is the question, in: Oncoplastic Breast Surgery, CRC Press, 2023, pp. 269–273.
[8] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[9] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, 2023, arXiv preprint arXiv:2304.07193.
[11] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W.L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, Z. Wang, ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools, 2024, arXiv:2406.12793.
[12] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, Pmlr, 2021, pp. 8821–8831.
[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[17] Y. Zhu, F. Cong, D. Zhang, W. Gong, Q. Lin, W. Feng, Y. Dong, J. Tang, Wingnn: Dynamic graph neural networks with random gradient aggregation window, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 3650–3662.
[18] Y. Zang, W. Li, J. Han, K. Zhou, C.C. Loy, Contextual object detection with multimodal large language models, 2023, arXiv preprint arXiv:2305.18279.
[19] Z. Chen, L. Luo, Y. Bie, H. Chen, Dia-LLaMA: Towards large language model-driven CT report generation, 2024, arXiv preprint arXiv:2403.16386.
[20] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, H. Tan, Lrm: Large reconstruction model for single image to 3d, 2023, arXiv preprint arXiv:2311.04400.
[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[22] S. Ren, Z. Wang, H. Zhu, J. Xiao, A. Yuille, C. Xie, Rejuvenating image-gpt as strong visual representation learners, in: Forty-First International Conference on Machine Learning, 2023.
[23] Z. Liu, Y. Li, P. Shu, A. Zhong, L. Yang, C. Ju, Z. Wu, C. Ma, J. Luo, C. Chen, et al., Radiology-llama2: Best-in-class large language model for radiology, 2023, arXiv preprint arXiv:2309.06419.
[24] Y. Feng, Z. Tian, Y. Zhu, Z. Han, H. Luo, G. Zhang, M. Song, CP-prompt: Composition-based cross-modal prompting for domain-incremental continual learning, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2729–2738.
[25] K. Shi, X. Sun, D. Wang, Y. Fu, G. Xu, Q. Li, LLaMA-E: Empowering E-commerce authoring with multi-aspect instruction following, 2023, arXiv preprint arXiv:2308.04913.
[26] N. Chen, P. Zelasko, L. Moro-Velázquez, J. Villalba, N. Dehak, Align-denoise: Single-pass non-autoregressive speech recognition, in: Interspeech, 2021, pp. 3770–3774.
[27] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, Flava: A foundational language and vision alignment model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15638–15650.
[28] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 1, (2) 2023, p. 3, arXiv preprint arXiv:2308.12966.
[29] A. Young, B. Chen, C. Li, C. Huang, G. Zhang, G. Zhang, H. Li, J. Zhu, J. Chen, J. Chang, et al., Yi: Open foundation models by 01. ai, 2024, arXiv preprint arXiv:2403.04652.
[30] J. Qi, M. Ding, W. Wang, Y. Bai, Q. Lv, W. Hong, B. Xu, L. Hou, J. Li, Y. Dong, et al., CogCoM: Train large vision-language models diving into details through chain of manipulations, 2024, arXiv preprint arXiv:2402.04236.
[31] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al., Cogvlm: Visual expert for pretrained language models, 2023, arXiv preprint arXiv:2311.03079.
[32] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, PMLR, 2023, pp. 19730–19742.
[33] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2024).
[34] H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, X. Zhang, Vary: Scaling up the vision vocabulary for large vision-language models, 2023, arXiv preprint arXiv:2312.06109.
[35] Q. Lin, Y. Zhu, X. Mei, L. Huang, J. Ma, K. He, Z. Peng, E. Cambria, M. Feng, Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey, Inf. Fusion (2024) 102795.
[36] Z. Deng, W. Gao, C. Chen, Z. Niu, Z. Gong, R. Zhang, Z. Cao, F. Li, Z. Ma, W. Wei, et al., OphGLM: An ophthalmology large language-and-vision assistant, Artif. Intell. Med. 157 (2024) 103001.
[37] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain? in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.
[38] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2023, arXiv preprint arXiv:2303.00915.
[39] T. Han, L.C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, K.K. Bressem, MedAlpaca–an open-source collection of medical conversational ai models and training data, 2023, arXiv preprint arXiv:2304.08247.
[40] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: Tuning llama model with chinese medical knowledge, 2023, arXiv preprint arXiv:2304.06975.
[41] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, Adv. Neural Inf. Process. Syst. 36 (2024).
[42] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, J. Teuwen, M. Broeders, G. Gennaro, P. Clauser, T.H. Helbich, M. Chevalier, T. Mertelmeier, Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study, Eur. Radiol. 29 (9) (2019) 4825–4832.
[43] K. Lång, M. Dustler, V. Dahlblom, A. Åkesson, I. Andersson, S. Zackrisson, Identifying normal mammograms in a large screening population using artificial intelligence, Eur. Radiol. (2020) 1–6.

[44] A. Yala, P.G. Mikhael, F. Strand, G. Lin, K. Smith, Y.-L. Wan, L. Lamb, K. Hughes, C. Lehman, R. Barzilay, Toward robust mammography-based models for breast cancer risk, Sci. Transl. Med. 13 (578) (2021).
[45] Z. Cao, Z. Yang, Y. Tang, Y. Zhang, M. Han, J. Xiao, J. Ma, P. Chang, Supervised contrastive pre-training formammographic triage screening models, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24, Springer, 2021, pp. 129–139.
[46] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.C. Corrado, A. Darzi, et al., International evaluation of an AI system for breast cancer screening, Nature 577 (7788) (2020) 89–94.
[47] M.V.S. de Cea, K. Diedrich, R. Bakalo, L. Ness, D. Richmond, Multi-task learning for detection and classification of cancer in screening mammography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 241–250.
[48] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al., Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3478–3488.
[49] H. Wu, F. Xiao, C. Liang, Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 417–434.
[50] P. Li, G. Liu, J. He, Z. Zhao, S. Zhong, Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 374–383.
[51] Y. Luo, W. Liu, T. Fang, Q. Song, X. Min, M. Wang, A. Li, Carl: Cross-aligned representation learning for multi-view lung cancer histology classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 358–367.
[52] Q. Jin, C. Zou, H. Cui, C. Sun, S.-W. Huang, Y.-J. Kuo, P. Xuan, L. Cao, R. Su, L. Wei, et al., Multi-modality contrastive learning for sarcopenia screening from hip X-rays and clinical information, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 85–94.
[53] Y. Zhong, M. Xu, K. Liang, K. Chen, M. Wu, Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest X-ray images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 724–733.
[54] H. Basak, Z. Yin, Pseudo-label guided contrastive learning for semi-supervised medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19786–19797.
[55] K. Kerlikowske, W. Zhu, R.A. Hubbard, B. Geller, K. Dittus, D. Braithwaite, K.J. Wernli, D.L. Miglioretti, E.S. O'Meara, B.C.S. Consortium, et al., Outcomes of screening mammography by frequency, breast density, and postmenopausal hormone therapy, JAMA Intern. Med. 173 (9) (2013) 807–816.
[56] M. Elshinawy, A. Badawy, W. Abdelmageed, M. Chouikha, Effect of breast density in selecting features for normal mammogram detection, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano To Macro, IEEE, 2011, pp. 141–147.
[57] C. D'Orsi, 2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System, American College of Radiology, ISBN: 9781559030168, 2014.
[58] S.G. Orel, N. Kay, C. Reynolds, D.C. Sullivan, BI-RADS categorization as a predictor of malignancy, Radiology 211 (3) (1999) 845–850.
[59] F.P. Kestelman, G.A.d. Souza, L.C. Thuler, G. Martins, V.A.R.d. Freitas, E.d. Canella, Breast imaging reporting and data system-BI-rads®: positive predictive value of categories 3, 4 and 5. a systematic literature review, Radiol. Bras. 40 (2007) 173–177.
[60] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1735–1742.
[61] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, G. Hu, Revisiting pre-trained models for Chinese natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Association for Computational Linguistics, Online, 2020, pp. 657–668, URL https://www.aclweb.org/anthology/2020.findings-emnlp.58.
[62] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, 2023, arXiv preprint arXiv:2306.13549.
[63] X. Chen, Y. Li, M. Hu, E. Salari, X. Chen, R.L. Qiu, B. Zheng, X. Yang, Mammo-CLIP: Leveraging contrastive language-image pre-training (CLIP) for enhanced breast cancer diagnosis with multi-view mammography, 2024, arXiv preprint arXiv:2404.15946.
[64] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.
[65] A.Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D.S. Chaplot, D.d.l. Casas, E.B. Hanna, F. Bressand, et al., Mixtral of experts, 2024, arXiv preprint arXiv:2401.04088.
[66] K.A. Lee, N. Talati, R. Oudsema, S. Steinberger, L.R. Margolies, BI-RADS 3: current and future use of probably benign, Curr. Radiol. Rep. 6 (2018) 1–15.
[67] S.G. Orel, N. Kay, C. Reynolds, D.C. Sullivan, BI-RADS categorization as a predictor of malignancy, Radiology 211 (3) (1999) 845–850.
[68] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[69] M. Howard, G. Agarwal, A. Lytwyn, Accuracy of self-reports of Pap and mammography screening compared to medical record: a meta-analysis, Cancer Causes & Control. 20 (2009) 1–13.
[70] R. Azevedo, Expertise in radiology: Accounting for the evidence and implications for instruction, 1999.
[71] K. Zhang, J. Yu, Z. Yan, Y. Liu, E. Adhikarla, S. Fu, X. Chen, C. Chen, Y. Zhou, X. Li, et al., Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks, 2023, arXiv preprint arXiv:2305.17100.
[72] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[73] K.A. Lee, N. Talati, R. Oudsema, S. Steinberger, L.R. Margolies, BI-RADS 3: current and future use of probably benign, Curr. Radiol. Rep. 6 (2018) 1–15.
[74] S. Raza, S.A. Chikarmane, S.S. Neilsen, L.M. Zorn, R.L. Birdwell, BI-RADS 3, 4, and 5 lesions: value of US in management—follow-up and outcome, Radiology 248 (3) (2008) 773–781.