Contents lists available at ScienceDirect

Medical Image Analysis

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

PneumoLLM: Harnessing the power of large language model for pneumoconiosis diagnosis

Meiyue Song[a,b,1], Jiarui Wang[c,1], Zhihua Yu[d,1], Jiaxin Wang[e,1], Le Yang[f,1], Yuting Lu[c], Baicun Li[g], Xue Wang[h,i], Xiaoxu Wang[c], Qinghua Huang[j], Zhijun Li[k,l], Nikolaos I. Kanellakis[m,n,o], Jiangfeng Liu[a,p,q,*], Jing Wang[a,b,*], Binglu Wang[c,*], Juntao Yang[a,p,q]

[a] *Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, 100005, China*
[b] *State Key Laboratory of Respiratory Health and Multimorbidity, Beijing, 100005, China*
[c] *School of Automation, Northwestern Polytechnical University, Shaanxi, Xi'an 710072, China*
[d] *Jinneng Holding Coal Industry Group Co. Ltd Occupational Disease Precaution Clinic, Shanxi, 037001, China*
[e] *School of Medicine, Tsinghua University, Beijing, 100084, China*
[f] *School of Electronics and Control Engineering, Chang'an University, Shaanxi, Xi'an 710064, China*
[g] *Center of Respiratory Medicine, China-Japan Friendship Hospital, National Center for Respiratory Medicine, Institute of Respiratory Medicine, Chinese Academy of Medical Sciences, National Clinical Research Center for Respiratory Diseases, Beijing, 100020, China*
[h] *Department of Respiratory, the Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, 150086, China*
[i] *Internal Medicine, Harbin Medical University, Harbin, Heilongjiang, 150081, China*
[j] *School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China*
[k] *Translational Research Center, Shanghai YangZhi Rehabilitation Hospital (Shanghai Sunshine Rehabilitation Center), Shanghai 201619, China*
[l] *School of Mechanical Engineering, Tongji University, Shanghai 201804, China*
[m] *Laboratory of Pleural and Lung Cancer Translational Research, CAMS Oxford Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK*
[n] *Oxford Centre for Respiratory Medicine, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK*
[o] *National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK*
[p] *Plastic Surgery Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100144, China*
[q] *State Key Laboratory of Common Mechanism Research for Major Diseases, Beijing, 100005, China*

ARTICLE INFO

Keywords:
Large language model
Medical image diagnosis
Foundational model

ABSTRACT

The conventional pretraining-and-finetuning paradigm, while effective for common diseases with ample data, faces challenges in diagnosing data-scarce occupational diseases like pneumoconiosis. Recently, large language models (LLMs) have exhibits unprecedented ability when conducting multiple tasks in dialogue, bringing opportunities to diagnosis. A common strategy might involve using adapter layers for vision–language alignment and diagnosis in a dialogic manner. Yet, this approach often requires optimization of extensive learnable parameters in the text branch and the dialogue head, potentially diminishing the LLMs' efficacy, especially with limited training data. In our work, we innovate by eliminating the text branch and substituting the dialogue head with a classification head. This approach presents a more effective method for harnessing LLMs in diagnosis with fewer learnable parameters. Furthermore, to balance the retention of detailed image information with progression towards accurate diagnosis, we introduce the contextual multi-token engine. This engine is specialized in adaptively generating diagnostic tokens. Additionally, we propose the information emitter module, which unidirectionally emits information from image tokens to diagnosis tokens. Comprehensive experiments validate the superiority of our methods.

1. Introduction

In the computer-aided diagnosis community, the processing and analysis prowess applied to medical data is pivotal. It facilitates the diagnosis of potential diseases and the prediction of future clinical outcomes. With the rapid evolution of deep learning theories, researchers have designed sophisticated network architectures (He et al., 2016; Dosovitskiy et al., 2020) and have curated extensive, high-quality datasets (Deng et al., 2009; Wang et al., 2017) to pretrain these

[*] Corresponding authors.
E-mail addresses: ljf@pumc.edu.cn (J. Liu), wangjing@ibms.pumc.edu.cn (J. Wang), wbl921129@gmail.com (B. Wang).

[1] Authors contributed equally.

https://doi.org/10.1016/j.media.2024.103248
Received 5 December 2023; Received in revised form 17 June 2024; Accepted 17 June 2024
Available online 20 June 2024

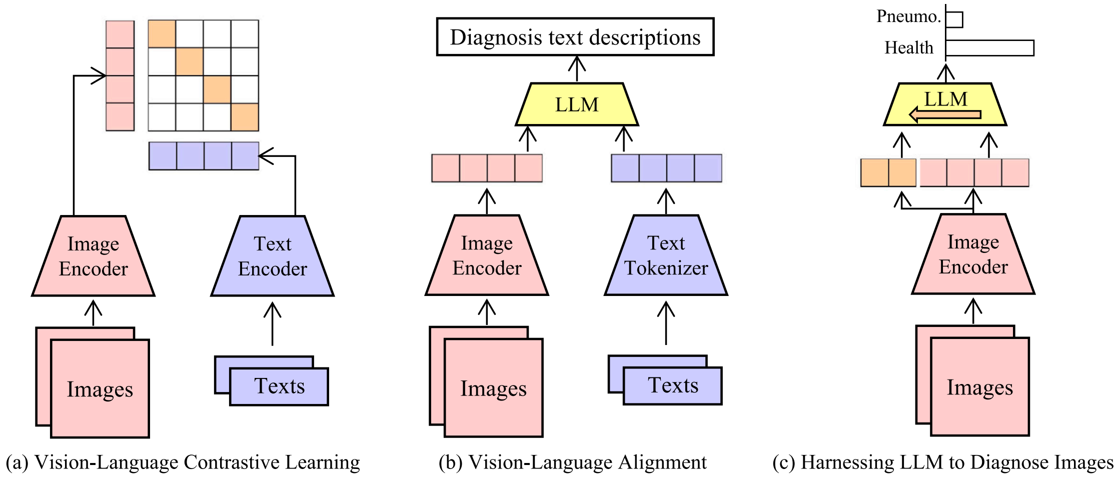


Image
Encoder
Text
Encoder
Images
Texts
(a) Vision-Language Contrastive Learning
Diagnosis text descriptions
LLM
Image
Encoder
Text
Tokenizer
Images
Texts
(b) Vision-Language Alignment
Pneumo.
Health
LLM
Image
Encoder
Images
(c) Harnessing LLM to Diagnose Images

Fig. 1. Representative pipelines to elicit knowledge from large models. (a) Traditional works conduct vision–language contrastive learning to align multimodal representations. (b) To utilize large language models, existing works transform images into visual tokens, and send visual tokens to LLM to generate text descriptions. (c) Our work harnesses LLM to diagnose medical images by proper designs, forming a simple and effective pipeline.

powerful networks. Pretraining strategies endow networks with valuable knowledge by optimizing weight distribution, which, in turn, equips researchers to further refine the model with labeled data targeted at diagnosing specific diseases. When the data is abundant and accurately labeled, this classical paradigm typically achieves commendable results, particularly with common ailments. An example is EchoNet-Dynamic (Ouyang et al., 2020), which has surpassed medical specialists in cardiac function assessment.

However, the landscape shifts when we delve into occupational diseases such as pneumoconiosis (Li et al., 2023b; Dong et al., 2022). Individuals subjected to prolonged exposure in dust-laden environments without personal protective equipments are susceptible to pulmonary fibrosis, a precursor to pneumoconiosis (Qi et al., 2021; Devnath et al., 2022). Areas with increased prevalence of pneumoconiosis often lack economical development, medical resources and infrastructure, and professional medical practitioners. Furthermore, there is a noticeable reticence towards disease screening and diagnosis, leading to an acute shortfall in clinical data for these diseases (Sun et al., 2023; Huang et al., 2023b). This paucity of data renders the conventional pretraining-and-finetuning strategy ineffective.

Addressing the diagnostic challenges posed by data-scarce occupational diseases requires an inventive approach. It involves tapping into the rich knowledge (Huang et al., 2023d) of foundational models and controlling the amount of learnable parameters to streamline the learning process. The advent of large language models (LLMs) (Kenton and Toutanova, 2019; Brown et al., 2020) has gained a bounty of knowledge from voluminous pretraining corpora, showcasing an impressive generalization capability for new tasks. The medical image diagnosis community has witnessed the emergence of foundational models (Zhang and Metaxas, 2024; Zhang et al., 2024), with significant strides made in pathology image analysis (Zhang et al., 2023a; Huang et al., 2023a) and medical image segmentation (Cheng et al., 2023; Lei et al., 2023; Wang et al., 2023a).

In the wake of such LLMs breakthroughs, concerted efforts continue to leverage knowledge from large-scale models to enhance image processing tasks, as depicted in Fig. 1. For instance, CLIP (Radford et al., 2021) embarks on vision–language contrastive learning to carefully align visual and language representations, as showcased in Fig. 1(a). The multimodal community, in turn, benefits from the integration of LLMs by interpreting visual tokens as a specialized form of language and devising adapters (Li et al., 2023c; Zhang et al., 2023b) to convert visual inputs into comprehensible representations, as shown in Fig. 1(b). These works often employ advanced techniques, such as instruction tuning (Stiennon et al., 2020; Liu et al., 2024), to yield fluent and varied narrative outputs. Meanwhile, some works begin to explore the interactive tool for interpreting the inner attention mechanisms of large vision–language models (Stan et al., 2024). The medical image diagnosis sector has also made notable advances by developing vision–language models (Wen et al., 2023; Yi et al., 2023) or by constructing medical foundational models from scratch (Moor et al., 2023). Xu et al. (2023b) propose a unified transformer model specifically designed for multi-modal clinical tasks by incorporating customized instruction tuning. Li et al. (2024) introduce conversational generative AI into the biomedicine domain and can follow open-ended instruction to assist with inquiries about a biomedical image.

Nonetheless, the application of these existing LLM-based pipelines to diagnose data-scarce occupational diseases poses several challenges. Firstly, the dependency on ample paired image-text data intensifies the complexity of data gathering, particularly when factoring in the constraints of patient privacy. Besides, processing text inputs through separate branches escalates computational demands substantially. Although the textual outputs are versatile, they may be unnecessarily complex for tasks that require simple binary outcomes, such as affirming the presence or absence of a specific disease.

Our approach diverges from existing methodologies by *eliminating the textual processing branch and directly harnessing LLMs to process images for the diagnosis of pneumoconiosis*, as shown in Fig. 1(c). We hypothesize that LLMs, after extensive corpus learning, are adept at selecting salient visual tokens and filtering out irrelevant ones, thereby benefiting the medical image diagnostic process. Fig. 2 presents the proposed PneumoLLM framework. We revise the language prediction head into a classification head, transitioning from dialogue-based outputs to succinct disease classification. After freezing parameters of both the vision encoder and the LLM, we integrate the adapter module and manage the learnable parameters effectively. We ascertain that eliciting diagnostic knowledge from LLMs hinges on balancing the preservation of comprehensive image details with the progression towards specific diagnostic task. To navigate this balance, we introduce the contextual multi-token engine that generates diagnostic tokens conditioned on image tokens. This ensures that the source image tokens retain all the pertinent image details. Subsequently, the information emitter module is engineered to unidirectionally emit information from source tokens to diagnosis tokens, thus steering the learning trajectory towards accurate diagnosis. All code is available at https://github.com/CodeMonsterPHD/PneumoLLM.

In brief, this work contributes to the field by:

- Charting new paradigm in applying LLMs to medical image analysis, especially for data-scarce occupational diseases, thereby simplifying the diagnostic process while yielding promising results.

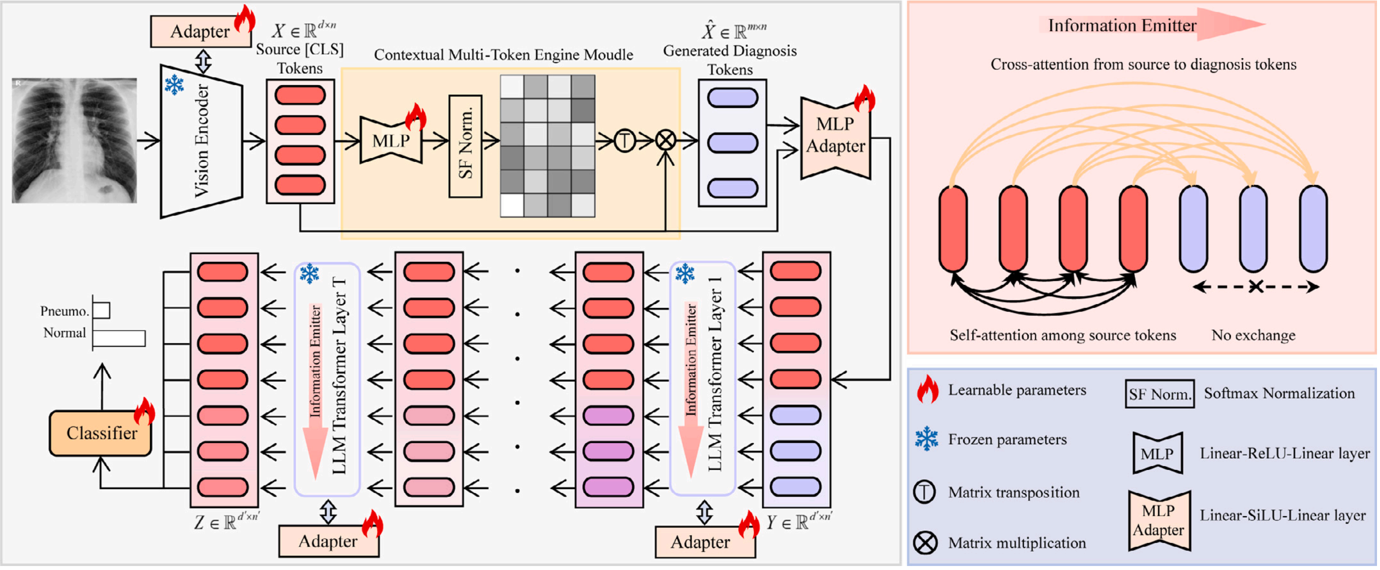

Adapter
$X \in \mathbb{R}^{d \times n}$
Source [CLS] Tokens
Contextual Multi-Token Engine Moudle
$\hat{X} \in \mathbb{R}^{m \times n}$
Generated Diagnosis Tokens
Vision Encoder
MLP
SF Norm.
T
MLP Adapter
Pneumo.
Normal
Classifier
Information Emitter
LLM Transformer Layer T
LLM Transformer Layer 1
$Z \in \mathbb{R}^{d' \times n'}$
$Y \in \mathbb{R}^{d' \times n'}$
Adapter
Adapter
Information Emitter
Cross-attention from source to diagnosis tokens
Self-attention among source tokens
No exchange
Learnable parameters
Frozen parameters
Matrix transposition
Matrix multiplication
SF Norm. Softmax Normalization
MLP Linear-ReLU-Linear layer
MLP Adapter Linear-SiLU-Linear layer

Fig. 2. Diagram of the proposed PneumoLLM. The vision encoder processes chest radiography and extracts source tokens. The contextual multi-token engine generates multiple diagnosis tokens conditioned on source tokens. To elicit in-depth knowledge from the LLM, we design the information emitter module within the LLM Transformer layers, enabling unidirectional information flow from source tokens to diagnosis tokens, preserving complete radiographic source details and aggregating critical diagnostic information.

- Designing the novel contextual multi-token engine and information emitter module to meticulously draw out knowledge from LLMs, achieving a harmonious balance between preserving image representations and harnessing LLMs diagnostic intelligence.
- Demonstrating the superiority of our method in diagnosing pneumoconiosis through extensive experiments and validating the effectiveness of each designed module.

2. Related work

2.1. Disease diagnosis based on X-ray

Recent advancements in deep learning have shown significant promise in the field of medical image diagnosis. Luo et al. (2022a) established a large-scale whole abdominal organ Dataset (WORD) for deep learning-based algorithm research and clinical application development. Luo et al. (2022b) developed a 3D sphere representation-based center-points matching detection network specifically for detecting pulmonary nodules in CT images. Wu et al. (2023a) proposed a pattern-aware transformer to achieve hierarchical pattern matching and propagation among sequential medical images. Ali et al. (2020) introduced a novel additive angular metric for few-shot classification of diverse endoscopy samples within a prototypical network framework. Kang et al. (2022) proposed a method for intra- and inter-task consistent learning, enhancing model predictions across various related tasks and addressing inconsistencies inherent in such tasks.

In the realm of anomaly detection, Li et al. (2023a) pioneered an unsupervised framework, SSL-AnoVAE, which leverages self-supervised learning (Huang et al., 2023c; Wang et al., 2023c) to provide fine-grained semantic analysis for anomalies in retinal images. Huang et al. (2022) introduced a transformer-based approach for classifying pneumoconiosis in 3D CT images, effectively combining intra-slice and inter-slice interaction information. Chen et al. (2023a) proposed a dynamic feature splicing strategy for few-shot diagnosis of rare diseases (e.g., hernia), employing similarity channel replacement at both low and high feature levels. Xing et al. (2023) presented a two-stage diagnostic framework involving multi-modal learning and cross-modal distillation, addressing challenges of limited dataset size and structural variations. Xu et al. (2023a) combined deep learning and machine learning methods for segmentation and feature extraction, mimicking the workflow of experienced radiologists. Chen et al. (2023b) introduced OLFG, an orthogonal latent space learning approach with feature weighting and graph learning for multimodal Alzheimer's Disease diagnosis. Ma et al. (2023) developed MGCA-RAFFNet, a multi-graph cross-attention-based network for brain disorder diagnosis, utilizing multi-template analysis. Fan et al. (2023) extended conventional siamese networks for low-shot learning, introducing a semi-supervised strategy that utilizes unlabeled data to enhance accuracy.

As for disease diagnosis based on X-ray images, Wang et al. (2020a) proposed COVID-Net, a ResNext50 network pre-trained on ImageNet and employing a lightweight PEPX design pattern. Zheng et al. (2019) restructured GoogLeNet using convolutional kernel decomposition. Wang et al. (2020b) utilized GoogleNet (Inception-v3) to detect pneumoconiosis. Devnath et al. (2021) employed two Convolutional Neural Network (CNN) models for feature extraction in pneumoconiosis CR images. Gao et al. (2021) developed a vision transformer based on attention models and DenseNet for COVID-19 classification from 2D slices of 3D CT images. Heidarian et al. (2021) proposed a CAE-Transformer framework for efficient classification of lung adenocarcinoma tumors using whole 3D CT images. The methods for diagnosing pneumoconiosis are summarized in Table 1. However, the reliance on data-driven deep learning methods necessitates ample training data, presenting challenges for occupational diseases like pneumoconiosis.

2.2. Foundational models and applications to diagnosis

The emergence of foundational models in the natural language domain, exemplified by the pre-training of Transformer (Vaswani et al., 2017) and BERT (Kenton and Toutanova, 2019), has demonstrated remarkable generalization abilities. Researchers have developed various parameter-efficient fine-tuning strategies, such as prefix tuning (Lester et al., 2021) and adapter methods (Houlsby et al., 2019), to leverage the potential of these models, often achieving competitive or superior performance in downstream tasks. The advent of CLIP, through its contrastive pre-training approach, has established a robust vision–language foundational model (Radford et al., 2021). This development has significantly advanced zero-shot and few-shot learning tasks, facilitated by innovative prompt tuning strategies (Zhou et al., 2022b) and adapter techniques (Hu et al., 2021).

Table 1
Existing diagnosis methods for pneumoconiosis.

Method	Advantages	Disadvantages
Huang et al. (2022)	Captured intra-slice dependencies and inter-slice information exchange.	Large number of model parameters.
Zheng et al. (2019)	Modeled features at different scales.	Complex network structure that is prone to overfitting on small datasets
Wang et al. (2020b)	Improved the network's ability to learn features at different scales and positions.	High training complexity and complex parameter configuration.
Devnath et al. (2021)	Used a multi-layer feature aggregation method to address the dust lung disease detection on a small dataset.	Sensitive to hyperparameter selection.

Building upon the aforementioned progress in vision–language models, BLIP-2 (Li et al., 2023c) integrates pre-existing vision and language models, freezing the original parameters while learning an additional transformation network, thereby generating strong vision–language representations. In the vision domain, recent advancements in foundational models have focused on providing general representations for a variety of downstream tasks (Oquab et al., 2023) and enhancing performance in open-world environments, including segmenting any object (Kirillov et al., 2023) and recognizing diverse entities (Zhang et al., 2023c). In the language domain, development efforts have led to the creation of large-scale foundational models, such as PaLM (Chowdhery et al., 2022) and ChatGPT (OpenAI, 2023a), which, being non-open-sourced, are accessible only through APIs. Conversely, other efforts have produced open-sourced models like LLaMA (Touvron et al., 2023a), opening new avenues for research.

The medical image diagnosis domain has also benefitted considerably from foundational models (Gao et al., 2023). For example, Med-PaLM (Singhal et al., 2023) and DoctorGLM (Xiong et al., 2023) infuse extensive medical knowledge into general foundational models. Similarly, MedCLIP (Wang et al., 2022) and CXR-CLIP (You et al., 2023) utilize X-ray images to pretrain foundational models specialized for disease diagnosis. Subsequent research has focused on exploring and harnessing the rich knowledge embedded in these models, developing disease-specific adaptations through methods like prompt-tuning (Zhang et al., 2023a), adaptation (Wang et al., 2023b), and continual learning (Yi et al., 2023). Additionally, efforts have been made to develop multimodal foundational models, such as PLIP (Huang et al., 2023a) and RadFM (Wu et al., 2023b), targeting a wide array of diagnostic tasks in a unified manner.

Despite the promising potential of foundational models, they typically require a substantial volume of paired image-text training data and often generate predictions in a dialogue format. In the context of pneumoconiosis diagnosis, the available data is limited to hundreds of images, and the annotations are classification labels rather than dialogue sentences. Therefore, this work represents an early exploration into harnessing the rich knowledge within foundational language models for direct application to image diagnosis tasks.

3. Method

3.1. Overview

The efficacy of computer-aided diagnosis systems is crucial in processing and analyzing medical data. However, these systems often face a significant shortfall in clinical data availability. Leveraging the rich knowledge reservoirs of foundational models is a promising strategy to address this data scarcity. Yet, the conventional pretraining-and-finetuning approach may compromise the representation capabilities of LLMs, due to substantial changes in their parameter spaces, leading to increased training time and memory overhead (Touvron et al., 2023a,b; OpenAI, 2023b).

To mitigate these challenges, we introduce PneumoLLM, an innovative LLM-based framework tailored for diagnosing pneumoconiosis using chest radiographs. PneumoLLM begins by processing chest radiographs (Fig. 2) through a vision encoder to extract informative source tokens, which are subsequently input into the LLMs to derive the final classification results. To ensure effective integration between the vision encoder and the LLMs, as well as to enhance the adaptability of the LLM to specific diagnosis task while preserving its original structure, we propose the integration of additional diagnosis tokens. To achieve this, we introduce two modules: contextual multi-token engine and information emitter modules. The former is responsible for generating the additional contextual diagnosis tokens, while the latter emits information from source tokens to additional contextual diagnosis tokens, preserving complete radiography source details and consolidating valuable diagnostic information. Additionally, to avoid disrupting the LLM's robust representation, we introduce adapter layers in both the vision encoder and the LLM model.

In detail, a chest radiography image, denoted as I_{img}, is processed through a vision encoder f_{vis}. We utilize the image encoder from the pretrained CLIP-ViT (Radford et al., 2021) to capitalize on its intrinsic alignment with language data, thereby facilitating the comprehension by subsequent LLMs. The visual features, represented as $X \in \mathbb{R}^{d \times n}$, are extracted using `[CLS]` tokens from every fourth layer of the ViT, where d is the number of extracted `[CLS]` tokens and n is the feature dimension of each token. Subsequently, X is processed by our contextual multi-token engine module, generating additional contextual diagnosis tokens $\hat{X} \in \mathbb{R}^{m \times n}$, where m is the number of newly generated tokens. The original `[CLS]` tokens X and the new contextual diagnosis tokens $\hat{X}$ are then concatenated and passed through a simple adapter MLP layer. This layer, comprising a simple two-layer bottleneck structure (Linear-SiLU-Linear) with the hidden layer dimension reduced to 128, transforms the amalgamated visual tokens $X' = [X, \hat{X}]$ into dimensions compatible with the LLM, resulting in $Y \in \mathbb{R}^{d' \times n'}$, where $d' = d + m$ is the total number of tokens resulting from the concatenation of X and $\hat{X}$, and n' is the feature dimension of each token in LLM.

The processed tokens Y are input into the pretrained LLM, sans its final classifier layer, to generate the final features $Z \in \mathbb{R}^{d' \times n'}$. These features are then fed into our disease classification network $h_{\varphi}(\cdot)$, designed for pneumoconiosis diagnosis, yielding the final classification logit scores $J \in \mathbb{R}^{c}$, where c denotes the number of classification categories. In line with the previous approaches (Luo et al., 2023b; Zhang et al., 2023b), we integrate the adapter layer $h_{\phi}(\cdot)$ to each multi-head attention module in both the vision encoder and LLM layers. During training, the learnable parameters in the adapter, the contextual multi-token engine module, and the disease classification network undergo training via a cross-entropy loss function, while the rest of the PneumoLLM parameters remain frozen.

3.2. Contextual multi-token engine

In the field of vision–language alignment, some data-efficient approaches like CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) advocate prompt engineering to achieve better performance than fixed hand-crafted prompts, which serves as an inspiration for our design. The promote engineering may also be useful in directly guiding the LLMs to further complement the information of visual tokens beyond mere vision–language alignment. However, CoOp's limitation lies in its uniform prompts for all samples, restricting instance flexibility during inference. CoCoOp extends this by learning a vision-conditional prompt token, moderately enhancing inference adaptability. Yet, this direct addition of fixed and flexible prompts might confuse LLMs and underutilize its potential. Our proposed contextual multi-token engine module aims generate new diagnosis tokens to seamlessly integrate and maximize the utility of information across diverse vision tokens. They will guide LLMs to generate more complementary features, leading to accurate diagnosis.

Specifically, as shown in the left top of Fig. 2, we employ a two-layer contextual multi-token MLP network (Linear-ReLU-Linear) denoted as h_θ^1, with the hidden layer reducing the input dimension by 12. This network, along with Softmax normalization, is utilized to generate the output contextual attention map $M_c \in \mathbb{R}^{d\times m}$ conditioned on the original [CLS] tokens $X \in \mathbb{R}^{d\times n}$, where m is a hyper-parameter representing the number of generated diagnosis tokens. Subsequently, we employ matrix multiplication to compute the output diagnosis tokens $\hat{X} \in \mathbb{R}^{m\times n}$. The entire process can be described in Eq. (1).

$$M_c = \sigma(h_\theta^1(X)),\ \hat{X} = M_c^T \cdot X \tag{1}$$

3.3. Information emitter module

After obtaining the original [CLS] tokens X and the new contextual diagnosis tokens $\hat{X}$, we concatenate and process them through a simple adapter MLP layer h_θ^2, aligning the combined tokens $X' = [X; \hat{X}]$ with LLM-compatible dimensions, resulting in $Y \in \mathbb{R}^{d'\times n'}$, where $d' = d + m$ is the total number of tokens resulting from the concatenation of X and $\hat{X}$, and n' is the feature dimension of each token in LLM. Subsequently, as shown in the right top of Fig. 2, we develop the information emitter module to preserve the original LLM's information interactions for the source [CLS] tokens, while allowing the newly generated context-diagnostic tokens to extract and repurpose this information, thereby fostering novel insights for the diagnosis task.

Specifically, we improve the attention mechanism in each ViT layer of the LLM, preventing [CLS] token features from being altered by contextual diagnosis token features. We define a self-attention mask $M \in \mathbb{R}^{d'\times d'}$ and configure its values as follows:

$$M_{i,j} = \begin{cases} -\infty, & \text{if } j > d \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Subsequently, this mask guides the multihead self-attention process in each ViT layer, as formulated below:

$$\begin{aligned} &Q = W_q Y,\ K = W_k Y,\ V = W_v Y \\ &\mathrm{Attn}\,(Q, K, V) = \mathrm{Softmax}\left(QK^T/\sqrt{d'} + M\right) \cdot V \end{aligned} \tag{3}$$

According to the mask definition in Eq. (2), the attention calculation can be further formulated as:

$$\begin{aligned} &\mathrm{Attn}_s\,(Q, K, V) = \mathrm{Softmax}\left(Q_s K_s^T/\sqrt{d'}\right) \cdot V_s, \\ &\mathrm{Attn}_c\,(Q, K, V) = \mathrm{Softmax}\left(Q_c K_s^T/\sqrt{d'}\right) \cdot V_s, \\ &\mathrm{Attn}\,(Q, K, V) = \begin{bmatrix} \mathrm{Attn}_s\,(Q, K, V) \\ \mathrm{Attn}_c\,(Q, K, V) \end{bmatrix} \end{aligned} \tag{4}$$

where $\mathrm{Attn}_s\,(Q, K, V)$ represents the self-attention fusion resulting from the information in the original [CLS] tokens. Q_s, K_s, and V_s represent the query, key, and value information extracted from the original [CLS] tokens. $\mathrm{Attn}_c\,(Q, K, V)$ denotes the cross-attention fusion results, indicating the newly generated context diagnostic tokens learning from the original [CLS] tokens. Q_c represents the diagnostic query tokens, receiving the emitted information from source tokens.

By this design, we can ensure that the newly generated context diagnostic tokens will not affect the self-attention process of the original [CLS] tokens, but absorb emitted information from them and supplement their own information through the setting of mask. Besides, it should be noted that there is no information interaction between the newly generated context diagnostic tokens to ensure the uniqueness and diversity of promotes information.

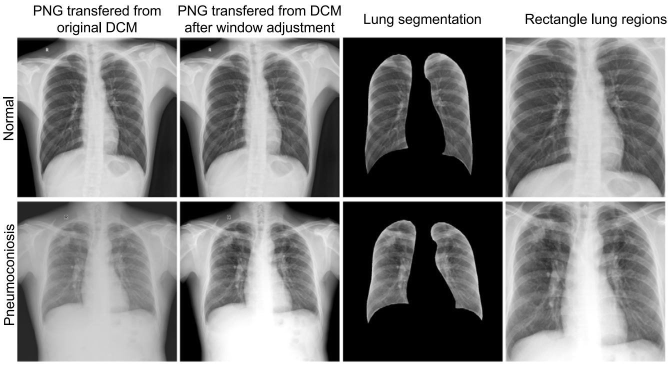

PNG transfered from original DCM
PNG transfered from DCM after window adjustment
Lung segmentation
Rectangle lung regions
Normal
Pneumoconiosis

Fig. 3. The illustration examples of dataset preprocessing: two examples labeled as "Normal" and "Pneumoconiosis". The window adjustment operation use the default window level and width (stored in the DICOM tags) to pre-process the original DICOM files. The segmentation results are obtained using the CheXmask pipeline, as proposed in the paper by Gaggion et al. (2023). The selection of the rectangular lung regions is based on the largest external rectangle of the segmentation results.

3.4. Network training

During the training phase, we focus on training the adapter layers $h_\phi(\cdot)$, the contextual multi-token engine network $h_\theta^1(\cdot)$, the simple adapter MLP layer $h_\theta^2(\cdot)$, and the disease classification network $h_\varphi(\cdot)$. Since the diagnosis of pneumoconiosis is a binary single-label task, we directly used binary cross-entropy loss for training:

$$\mathcal{L}_{\text{bce}}\left(R, \hat{R}; \phi, \theta_1, \theta_2, \varphi\right) = -R\log(\hat{R}) + (1 - R)\log(1 - \hat{R}) \tag{5}$$

where $\mathcal{L}_{\text{bce}}$ denotes the binary cross-entropy loss function, R represents the ground-truth pneumoconiosis diagnose labels, $\hat{R}$ represents the predicted results by our complete PneumoLLM framework, and $\phi, \theta_1, \theta_2, \varphi$ are the learnable parameters to be optimized.

Notably, when extending PneumoLLM to a multi-category classification task, the loss function may need to be replaced with cross-entropy loss. Furthermore, for multi-label classification, where a sample can belong to multiple categories simultaneously or none at all, the loss function might require multiple binary cross-entropy losses or other suitable losses for multi-label problems. After adjusting the loss function and the number of categories in the final classification linear layer, our PneumoLLM can also be directly applied to these tasks.

4. Experiments

4.1. Experimental setups

Dataset acquisition and splits. In this study, we utilized the posterior–anterior chest radiograph database from Jinneng Holding Coal Industry Group Co. Ltd Occupational Disease Precaution Clinic, comprising 630 chest radiographs in DICOM format, including 401 pneumoconiosis cases.

To ensure a balanced ratio of pneumoconiosis and normal samples in training and testing, we conduct five-fold cross-validation for experiments. Subsequently, we merge them into five distinct randomized datasets by patients (Datasets 1–5), and report the average performance of five experiments. Each dataset contains approximately 504 training and 126 testing radiographs, with pneumoconiosis cases constituting about 63% of each dataset.

Dataset pre-processing. In the preprocessing phase, we first use the default window level and width (stored in the DICOM tags) to pre-process the original DICOM files, as in the approach described by Wang et al. (2023b). This step can adjust the contrast and brightness level of the chest radiographs, making the anatomical structure or lesion of interest easier to distinguish and observe. After the window adjustment,

Table 2
Comparison results with recent prestigious methods on the pneumoconiosis dataset.

Method	L-para.(M)	Sens. (%)	Spec. (%)	Acc. (%)	AUC (%)	AVG (%)
ResNet (He et al., 2016)	25.56	76.55	54.49	68.57	71.11	67.68
ViT (Dosovitskiy et al., 2020)	88.30	73.56	51.96	65.72	64.23	63.87
Swin Transformer (Liu et al., 2021)	28.29	74.08	52.77	66.34	67.17	65.09
Conformer (Peng et al., 2021)	23.25	70.81	59.88	66.82	70.10	66.90
ConvNeXt (Liu et al., 2022)	88.59	69.35	62.84	66.99	67.64	66.70
DINOv2 (Oquab et al., 2023)	22.06	75.56	57.23	68.89	70.40	68.02
VAPFormer (Kang et al., 2023)	1.20	63.04	59.26	61.15	63.39	61.71
PneumoLLM	2.70	**80.54**	**67.66**	**75.87**	**78.98**	**75.76**

Table 3
Comparison results with recent LLM-based methods on the pneumoconiosis dataset.

Method	L-para.(M)	Sens. (%)	Spec. (%)	Acc. (%)	AUC (%)	AVG (%)
Zero-Shot CLIP (Radford et al., 2021)	—-	**98.75**	0	62.70	57.62	54.77
Linear Probe CLIP (Radford et al., 2021)	0.0008	53.71	75.08	67.31	27.96	56.02
CoOp (Zhou et al., 2022b)	0.003	75.05	51.10	66.35	67.90	65.10
CoCoOp (Zhou et al., 2022a)	0.078	74.58	60.67	69.53	71.14	68.98
LaVIN (Luo et al., 2023b)	3.77	87.80	47.52	73.33	71.82	70.12
BLIP2 (Li et al., 2023c)	107.13	69.88	**69.47**	69.70	77.90	71.12
PneumoLLM	2.70	80.54	67.66	**75.87**	**78.98**	**75.76**

we use the pyplot.imsave function to convert the UInt16 DICOM format into the UInt8 PNG format. Next, we employ the CheXmask pipeline, introduced by Gaggion et al. (2023), for lung segmentation. Based on the lung segmentation results, we use the maximum external rectangle extraction technique to isolate the rectangular lung regions from the original chest radiographs. Finally, we resize these rectangular lung regions into a uniform size of 224 × 224 pixels for further analyses. Fig. 3 displays representative chest radiographs, showcasing both categories (e.g., pneumoconiosis and normal), the comparisons before and after the window adjustment operation, their corresponding lung segmentation results, and the extracted rectangular lung regions.

Evaluation metrics. Following Chen et al. (2023a) and Qu et al. (2023), we adapt Accuracy (Acc.), Sensitivity (Sens.), Specificity (Spec.), and Area Under the Curve (AUC) for quantitative analysis to comprehensively evaluate the performance of pneumoconiosis diagnosis. To facilitate analysis and comparison, we calculate the average of these metrics to assess the overall performance.

Implementation details. Consistent with (Luo et al., 2023b), we employ the ViT-L/14 (Dosovitskiy et al., 2020) of the pre-trained CLIP (Radford et al., 2021) as the vision encoder. We extract visual features as six `[CLS]` tokens from every fourth layer of ViT-L/14. The LLM utilize the LLaMA-7B (Touvron et al., 2023a) model. We set the visual neck dimension to 128 and the adapter dimension to 8. The adapter layer used in vision encoder and LLM is RepAdapter (Luo et al., 2023a). We employ AdamW (Loshchilov and Hutter, 2017) as the optimizer, training the model for 100 epochs with a cosine decay learning rate schedule. The batch size, learning rate, warmup epochs, and weight decay are set to 16, $3e^{-4}$, 2 and 0.02, respectively. Under this setting, due to the usage of frozen models and small trainable parameters, PneumoLLM fine-tuning is computationally efficient. For example, using a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory, PneumoLLM requires less than 25 min for fine-tuning 100 epochs on a small-size training dataset of 504 chest radiographs.

4.2. Comparison experiments

To evaluate our proposed PneumoLLM, we compare it against established natural image classification methods, including ResNet (Xie et al., 2017), ViT (Dosovitskiy et al., 2020), Swin Transformer (Liu et al., 2021), Conformer (Peng et al., 2021), ConvNeXt (Liu et al., 2022), and DINOv2 (Oquab et al., 2023), as well as the advanced medical image diagnosis method VAPFormer (Kang et al., 2023). All comparison models are implemented based on open-source configurations.

Pneumoconiosis diagnosis comparisons with recent prestigious methods. Table 2 details the quantitative performance metrics of our PneumoLLM against the prestigious image classification methods which are pretrained on ImageNet-1k (Deng et al., 2009) or LVD-142M (Oquab et al., 2023). Notably, among these prestigious image classification methods, models with inherent inductive biases like ResNet, Conformer, and ConvNeXt outperform transformer models like ViT and Swin Transformer in our limited Pneumoconiosis diagnosis dataset, while DINOv2 excels due to extensive pretraining data and prior information. Conversely, the medical image diagnosis method VAPFormer with prompts exhibits the least efficacy, which illustrates the big diagnostic challenges posed by data-scarce occupational disease. Our PneumoLLM, in contrast, shows promising results by harnessing the power of LLM requiring relatively fewer trainable parameters. In addition, we can observe that all the methods in Table 2 have significantly higher sensitivity and relatively lower specificity. The occurrence of this phenomenon is attributed to the unbalanced data distribution resulting from 401 out of the total 630 images in the dataset. Besides, we show some visual chest radiographs and performance comparisons by different algorithms in Fig. 4. When compared with different algorithms to diagnose pneumoconiosis, PneumoLLM showed higher confidence in accurate predictions.

Pneumoconiosis diagnosis comparisons with recent foundational models. Table 3 represents comparisons with recent existing foundational models, including Zero-Shot CLIP (Radford et al., 2021), Linear Probe CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a), representing vision–language contrastive learning methods (Fig. 1(a)), and BLIP2 (Li et al., 2023c) and LaVIN (Luo et al., 2023b), representing vision–language alignment-based methods (Fig. 1(b)). Notably, for the vision–language contrastive learning methods, we set the text labels corresponding to normal and pneumoconiosis chest radiographs as ‘Normal chest radiograph’ and ‘Pneumoconiosis chest radiograph’. And for the vision–language alignment-based methods, we treat the pneumoconiosis diagnosis as a question-answering task with a coincident question input: “Does the patient have pneumoconiosis?”. We create ‘Yes’ or ‘No’ text answer labels to ensure consistency with the traditional visual question answering experiment setting.

As shown in Table 3, directly using Zero-Shot CLIP on pneumoconiosis diagnosis tends to classify all chest radiographs as pneumoconiosis. This indicates that without pretraining, the vision features and language features extracted by CLIP are not discriminative due to the domain gap between natural images and medical chest radiographs. Even with a simple linear probe classifier on CLIP, the performance

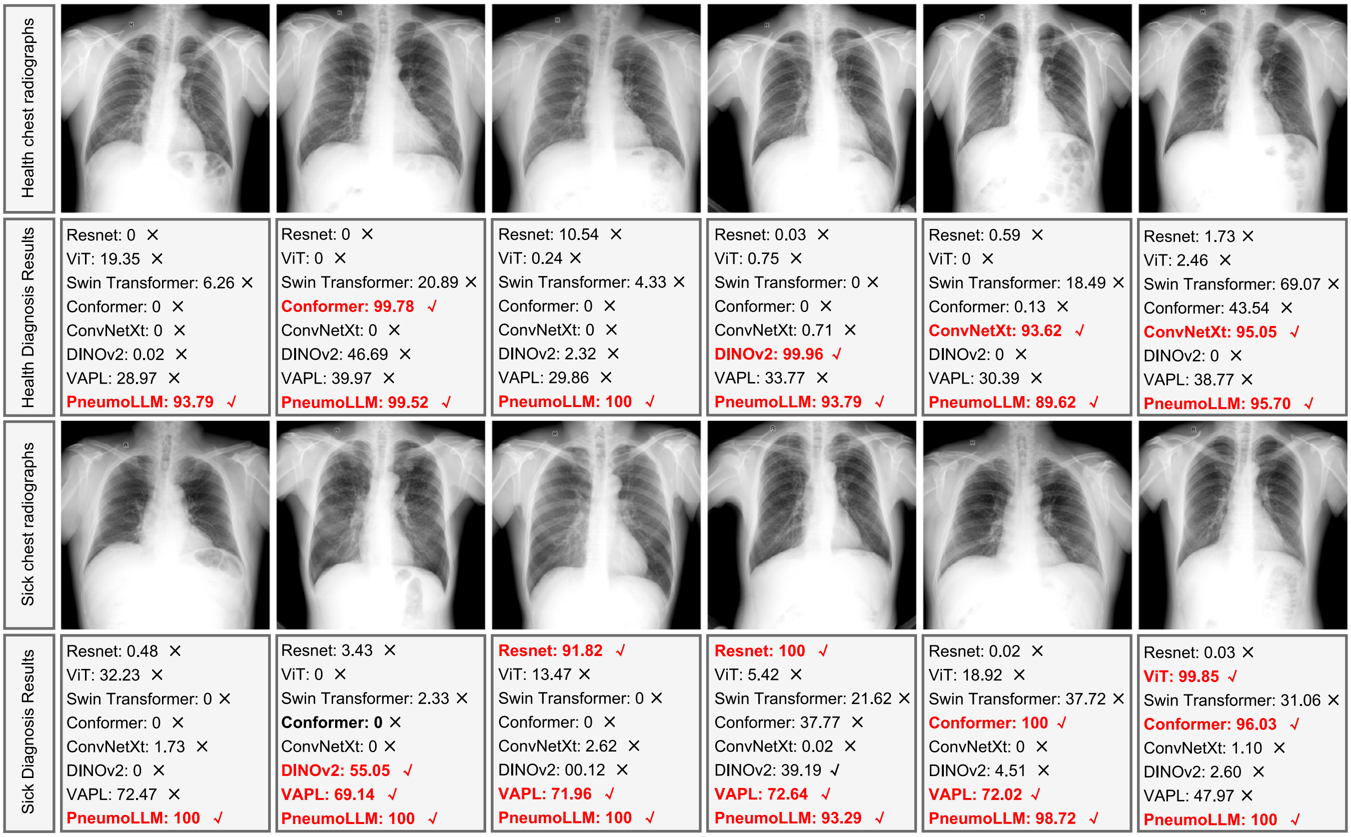


Fig. 4. Pneumoconiosis diagnosis results comparison with recent prestigious methods. The correct diagnosis results are highlighted in red.

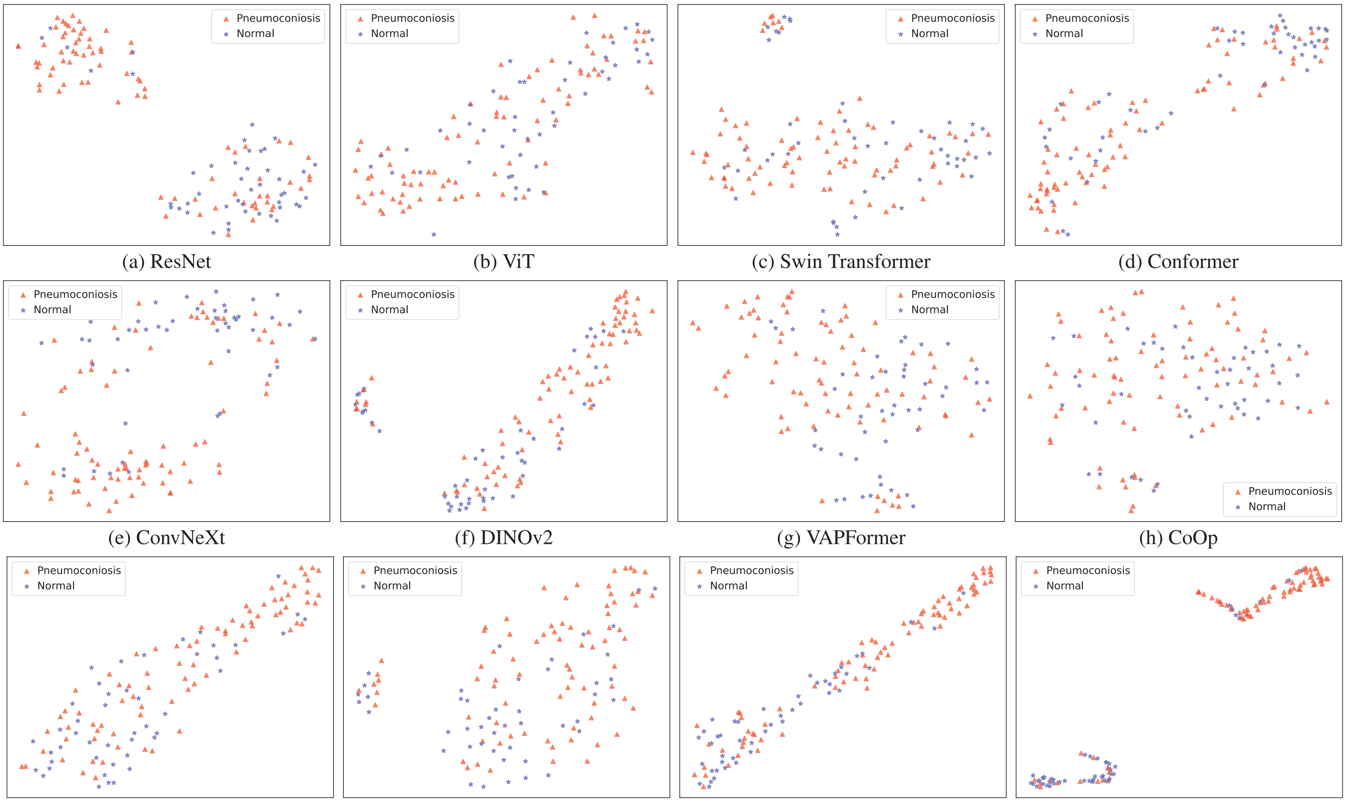
Table 4
Analysis of LLaMA-7B foundational model in pneumoconiosis diagnosis.

Models	L-para.(M)	Mem. (G)	Sens. (%)	Spec. (%)	Acc. (%)	AUC (%)	AVG (%)
PneumoLLM w/o LLaMA	0.52	4.15	77.77	62.00	72.06	75.81	71.91
PneumoLLM	2.70	15.48	**80.54**	**67.66**	**75.87**	**78.98**	**75.76** (+3.85)

remains undesirable, as indicated by the very low AUC evaluation metric value. This further highlights the inability of frozen CLIP trained on natural images to learn a valid linear association with the target diagnosis label. In contrast, using simple context prompt learning strategies, like CoOp and CoCoOp, can partly alleviate this problem and greatly improve the performance of pneumoconiosis diagnosis. However, these strategies only affect the language feature representation in CLIP, and their performance improvement ability is limited, suggesting the need for further adaptation. Furthermore, the vision–language alignment-based methods, LaVIN and BLIP2, which use LLM to understand the image tokens extracted from the image encoder and text instruction tokens, perform better than CoOp and CoCoOp. This illustrates that LLM does have the ability to improve the generalization of vision representation. However, while the Qformer designed in BLIP2 performs better in mitigating this gap compared to the adapter usage in LaVIN, the relatively large number of learnable parameters and complex pretraining strategies makes it less efficient in harnessing LLM for the pneumoconiosis diagnostic task. There is still plenty of room to explore improving the ability of LLM to understand the image tokens extracted from the image encoder and perform better with small computing cost. In light of these challenges, our proposed PneumoLLM outperforms all of these methods. By directly eliminating the text branch and substituting the dialogue head with a classification head, along with the subtle contextual multi-token engine and information emitter module design, our method achieves a harmonious balance between preserving image representations and harnessing LLM's diagnostic intelligence.

Qualitative comparisons with t-SNE visualization. To evaluate feature representation quality, we employ t-SNE (Van der Maaten and Hinton, 2008) method as qualitative comparisons to project high-level feature representations onto a 2D plane. We extracted the last-layer features and reshaped them into NCD-dimensional features, where N represents the number of samples, C denotes the channel dimension, and D represents the feature dimension. We then computed the mean along the channel dimension to obtain ND-dimensional features. This averaging operation collapses the channel information and generates a reduced representation for each sample. For CoOp and CoCoOp, since the decision criterion is based on the similarity comparison between visual features and text features of all categories, we concatenated all visual features and text features along the channel dimension and flattened them into ND-dimensional features. We then employed the t-SNE algorithm to project the ND-dimensional features onto a two-dimensional space for visualizing high-dimensional feature spaces. The final t-SNE visualization results are shown in Fig. 5.

Transformer-based architectures, like ViT, Swin Transformer, and LLM-based methods, like CoOp, do not perform optimally with limited datasets, exhibiting a less distinct separation between classes. In contrast, traditional convolutional networks like ResNet, along with hybrid models like Conformer and ConvNeXt, achieve better class discriminability. However, these models also present significant variance within class clusters. Notably, our proposed PneumoLLM model demonstrates a markedly superior clustering effect, with tightly grouped intra-class data points and stark demarcations between different classes. This compact and distinct representation suggests a more robust feature



(a) ResNet (b) ViT (c) Swin Transformer (d) Conformer

(e) ConvNeXt (f) DINOv2 (g) VAPFormer (h) CoOp

(i) CoCoOp (j) LaVIN (k) BLIP2 (l) Our PneumoLLM

Fig. 5. The t-SNE visualization of feature representation obtained by different networks in comparison experiment.

Table 5
Ablation study on eliminating the textual processing branch in LLM.

Settings	L-para.(M)	Mem.(G)	Sens. (%)	Spec. (%)	Acc. (%)	AUC (%)	AVG (%)
LaVIN	3.77	20.25	**87.80**	47.52	**73.33**	71.82	70.12
Simplified PneumoLLM	2.69	15.32	71.56	**71.67**	71.58	**75.21**	**72.50** (+2.38)

extraction capability. These visual insights are in concordance with our quantitative analyses, further corroborating the superior performance of PneumoLLM compared to conventional methodologies.

4.3. Ablation experiments

In our research, we perform a comprehensive ablation analysis of PneumoLLM, modifying one component at a time. Our PneumoLLM mainly comprises the utilization of LLM, contextual multi-token engine, and information emitter design. Initially, in Section 4.3.1, we analyze the model capacity, focusing on the essential role of the foundational LLaMA model, and the design of eliminating the textual processing branch by directly harnessing LLaMA to process visual features from the vision encoder. Subsequently, we evaluate the impact of various PneumoLLM components in Section 4.3.2, including the adapter, the contextual multi-token engine, and the information emitter module.

4.3.1. Model capacity

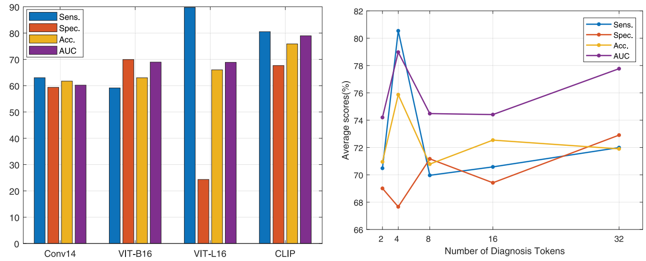
Ablation study on llama utilization. In contrast to existing classification vision models, our PneumoLLM incorporates the LLaMA-7B model, harnessing its extensive knowledge and prior information to improve pneumoconiosis diagnosis. To validate this approach, we perform an ablation study examining the effects of LLaMA integration, detailed in Table 4. Our default PneumoLLM configuration integrates the ViT-L/14+LLaMA architecture with additional learnable adapter layers, contextual multi-token engine network, and classification network. Without LLaMA, PneumoLLM solely utilizes the ViT-L/14 architecture with adapter layers and classification network. As indicated in Table 4, the utilization of LLaMA demonstrates superior learning efficiency on small-size pneumoconiosis datasets, achieving improvements across all diagnosis metrics. However, this improvement comes at the cost of increased memory due to the LLM's inherent model size.

Ablation on eliminating the textual processing branch in LLM. In contrast to LaVIN (Luo et al., 2023b), which treats image classification as a question-answering task and requires both question text and image as input to the LLM, we argue that the question text inputs are redundant and unnecessary for the vision classification task. Instead, we propose that directly inputting the [CLS] tokens obtained from the pre-trained CLIP into LLaMA is sufficient to enhance pneumoconiosis diagnosis. To validate this assumption, we conduct an ablation study comparing our PneumoLLM model with image-only input (Fig. 1(c)) against LaVIN's dual image-question input setup (Fig. 1(b)), where the setup of LaVIN is consistent with the one described in Table 3. For fairness, we build a simplified PneumoLLM by removing the contextual multi-token engine and information emitter module. The ablation results, presented in Table 5, demonstrate that our simplified PneumoLLM outperforms LaVIN in terms of both efficiency and efficacy, supporting our hypothesis that the inclusion of question text input is unnecessary for pneumoconiosis diagnosis.

To further assess the impact of various vision encoders and the necessity of the usage of pre-trained CLIP network, we compare various vision encoders, including a standard 14 × 14 convolutional layer, ViT-B/16 and ViT-L/16 pre-trained on ImageNet-21k, and ViT-L/14 from

Table 6
Ablation study on various PneumoLLM components.

Baseline	Adapter	CoOp	CoCoOp	Contextual multi-token engine	Information emitter	Sens. (%)	Spec. (%)	Acc. (%)	AUC (%)	AVG (%)
✓						78.07	62.77	72.54	74.52	71.98
✓	✓					71.56	**71.67**	71.58	75.21	72.50
✓	✓	✓				73.54	68.56	71.74	75.56	72.35
✓	✓		✓			72.57	70.31	71.75	75.80	72.61
✓	✓			✓		78.29	66.82	74.13	76.42	73.92
✓	✓			✓	✓	**80.54**	67.66	**75.87**	**78.98**	**75.76**



(a) Different vision encoder networks (b) Number of generated diagnosis tokens

Fig. 6. Illustration on various vision encoder networks and the number of generated diagnosis tokens. Please zoom in for the best view.

pre-trained CLIP (Radford et al., 2021). Results in Fig. 6(a) indicate the superior performance of the ViT-L/14 from pre-trained CLIP. Notably, ViT-L/16 exhibited high sensitivity but low specificity, suggesting a tendency to over-diagnose. This ablation study highlights the significance of choosing appropriate pre-trained vision encoder for optimal integration with LLM.

4.3.2. Ablation on various pneumollm components

Due to the limited availability of pneumoconiosis patient data and patient privacy concerns, acquiring extensive pneumoconiosis datasets is challenging. Fine-tuning our entire PneumoLLM on the small-size pneumoconiosis dataset can make drastic changes in the LLM's parameter spaces, leading to increased training time and huge memory requirements. In this scenario, the usage of lightweight adapters, our proposed contextual multi-token engine and information emitter module prove crucial in adapting LLM efficiently to pneumoconiosis diagnosis. To validate this, we conduct comprehensive ablation studies in Table 6. In this table, we use the simplified PneumoLLM in Table 5 without adapters as our baseline for comparing various component configurations.

Ablation on adapter usage. We conduct an ablation study to measure the effect of the usage of adapters. The results, as presented in the first two rows of Table 6, indicate that the incorporation of adapters significantly enhances diagnostic performance in our small-size pneumoconiosis dataset.

Ablation on the contextual multi-token engine. As discussed in Section 3.2, CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) are two recent typical prompt engineering methods. To highlight the effectiveness of our proposed contextual multi-token engine design, we replace our contextual multi-token engine with the promote design in CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) while keeping other components the same. The ablation results are in Table 6. While CoOp and CoCoOp settings show limited improvement, our multi-token engine yields a significant performance increase (+1.98%). CoOp's uniform prompts across samples hinder its adaptability, leading to a slight decrease in average performance. CoCoOp, though offering improved flexibility with vision-conditional prompts, faces limitations due to the mixed use of fixed and flexible prompts, yielding minimal gains. In contrast, our multi-token engine, designed to optimize the synergy between vision tokens and LLM, significantly enhances pneumoconiosis diagnosis.

Ablation on the information emitter module. Another ablation study focuses on our proposed information emitter module. Results in Table 6 demonstrate that combining the multi-token engine with the information emitter module achieves the best performance in pneumoconiosis diagnosis. This module successfully maintains the integrity of the original LLM's information processing while emitting valuable information to context-diagnostic tokens, fostering enhanced diagnostic insights.

Additionally, the optimal number of generated diagnosis tokens is determined based on performance analysis (Fig. 6(b)), leading to the selection of four as the ideal number.

5. Conclusion

In this paper, we introduce PneumoLLM, a pioneering approach utilizing large language models for streamlined diagnostic processes in medical imaging. By discarding the text branch and transforming the dialogue head into a classification head, PneumoLLM simplifies the workflow for eliciting knowledge from LLMs. This innovation proves particular effectiveness when only classification labels are available for training, rather than extensive descriptive sentences. The streamlined process also significantly reduces the optimization space, facilitating learning with limited training data. Ablation studies further underscore the necessity and effectiveness of the proposed modules, especially in maintaining the integrity of source image details while advancing towards accurate diagnostic outcomes.

Still, our study has some limitations. First, our current work primarily explored how effectively leveraging LLMs can improve performance on the small pneumoconiosis diagnosis dataset. However, when dealing with multi-category and multi-label tasks, severe long-tailed classification problems may disrupt the model's learning process. Facing these more challenging complex scenarios, we need to consider more sophisticated improvements related to the realistic attributes of categories, such as incorporating causal multi-relationship graph designs when transferring source [CLS] tokens into diagnosis tokens in the contextual multi-token engine module. Additionally, the adapter layers and the final disease classification network may also require modifications to handle the increased complexity and diversity of diagnostic categories effectively. Looking ahead, we plan to expand PneumoLLM's application to more imaging modalities beyond chest radiography, e.g., CT and MRI scans, aiming to broaden its diagnostic capabilities across a spectrum of diseases. Second, while removing the text branch is successful in our current pneumoconiosis classification task which only focuses on the single vision modality, such an operation limits its application to more complex text-based multi-tasks in the medical domain, like multi-modality joint diagnosis and open-ended conversational generation (Li et al., 2024). Therefore, in future research, we need to continue to find the better trade-off between computational complexity and the interactive fusion of features explored by different modalities. These future endeavors could enhance the capabilities of automated diagnostic systems, paving the way for more practical medical imaging analyses.

CRediT authorship contribution statement

Meiyue Song: Writing – review & editing, Formal analysis, Data curation, Conceptualization. **Jiarui Wang:** Writing – original draft, Visualization, Methodology, Investigation. **Zhihua Yu:** Data curation, Conceptualization. **Jiaxin Wang:** Investigation, Formal analysis, Data curation. **Le Yang:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Yuting Lu:** Writing – original draft, Visualization, Methodology. **Baicun Li:** Formal analysis, Data curation. **Xue Wang:** Investigation, Formal analysis. **Xiaoxu Wang:** Supervision, Methodology. **Qinghua Huang:** Writing – review & editing, Formal analysis. **Zhijun Li:** Supervision, Methodology. **Nikolaos I. Kanellakis:** Writing – review & editing, Formal analysis. **Jiangfeng Liu:** Conceptualization. **Jing Wang:** Supervision, Project administration, Data curation. **Binglu Wang:** Supervision, Project administration, Funding acquisition. **Juntao Yang:** Validation, Supervision, Resources, Funding acquisition.

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.

Data availability

We have shared the github link with our code and pretrained model weight in the manuscript. But the authors are not promising to share the data.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT4.0 in order to improve readability and language of the work. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work was supported in part by grants from the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences, China CIFMS2021-I2M-1-044, 2021-I2M-1-049, and the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences, 2022-RC310-06, and the NSFC, China 62071382, 82272020, and the Basic and Applied Basic Research Foundation of Guangdong Province 2024A1515012388. Juntao Yang is a senior author to this work.

References

Ali, S., Bhattarai, B., Kim, T.K., Rittscher, J., 2020. Additive angular margin for few shot learning to classify clinical endoscopy images. In: Proceedings of the International Workshop on Machine Learning in Medical Imaging. Springer, pp. 494–503.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Chen, Y., Guo, X., Pan, Y., Xia, Y., Yuan, Y., 2023a. Dynamic feature splicing for few-shot rare disease diagnosis. Med. Image Anal. 90, 102959.

Chen, Z., Liu, Y., Zhang, Y., Li, Q., Initiative, A.D.N., et al., 2023b. Orthogonal latent space learning with feature weighting and graph learning for multimodal alzheimer's disease diagnosis. Med. Image Anal. 84, 102698.

Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K., 2023. Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. PaLM: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Devnath, L., Fan, Z., Luo, S., Summons, P., Wang, D., 2022. Detection and visualisation of pneumoconiosis using an ensemble of multi-dimensional deep features learned from chest X-rays. Int. J. Environ. Res. Public Health 19 (18), 11193.

Devnath, L., Luo, S., Summons, P., Wang, D., 2021. Automated detection of pneumoconiosis with multilevel deep features learned from chest X-Ray radiographs. Comput. Biol. Med. 129, 104125.

Dong, H., Zhu, B., Zhang, X., Kong, X., 2022. Use data augmentation for a deep learning classification model with chest X-ray clinical imaging featuring coal workers' pneumoconiosis. BMC Pulm. Med. 22 (1), 1–14.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–12.

Fan, R., Bowd, C., Brye, N., Christopher, M., Weinreb, R.N., Kriegman, D.J., Zangwill, L.M., 2023. One-vote veto: Semi-supervised learning for low-shot glaucoma diagnosis. IEEE Trans. Med. Imaging.

Gaggion, N., Mosquera, C., Mansilla, L., Aineseder, M., Milone, D.H., Ferrante, E., 2023. CheXmask: A large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. arXiv preprint arXiv:2307.03293.

Gao, Y., Li, Z., Liu, D., Zhou, M., Zhang, S., Meta, D.N., 2023. Training like a medical resident: Universal medical image segmentation via context prior learning. arXiv preprint arXiv:2306.02416.

Gao, X., Qian, Y., Gao, A., 2021. COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models. arXiv preprint arXiv:2107.01682.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heidarian, S., Afshar, P., Oikonomou, A., Plataniotis, K.N., Mohammadi, A., 2021. Cae-transformer: Transformer-based model to predict invasiveness of lung adenocarcinoma subsolid nodules from non-thin section 3d ct scans. arXiv preprint arXiv:2110.08721.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning. PMLR, pp. 2790–2799.

Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al., 2021. LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations. pp. 1–16.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023a. A visual–language foundation model for pathology image analysis using medical twitter. Nature Med. 1–10.

Huang, B., Liao, G., Lee, P.M.Y., Chan, C.K., Tai, L.b., Tsang, C.Y.J., Leung, C.C., Tse, L.A., 2023b. Association of circadian rhythm with mild cognitive impairment among male pneumoconiosis workers in Hong Kong: A cross-sectional study. Sci. Rep. 13 (1), 1650.

Huang, Y., Si, Y., Hu, B., Zhang, Y., Wu, S., Wu, D., Wang, Q., 2022. Transformer-based factorized encoder for classification of pneumoconiosis on 3D CT images. Comput. Biol. Med. 150, 106137.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023c. STU-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716.

Huang, Q., Wang, D., Lu, Z., Zhou, S., Li, J., Liu, L., Chang, C., 2023d. A novel image-to-knowledge inference approach for automatically diagnosing tumors. Expert Syst. Appl. 229, 120450.

Kang, L., Gong, H., Wan, X., Li, H., 2023. Visual-attribute prompt learning for progressive mild cognitive impairment prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 547–557.

Kang, Q., Lao, Q., Li, Y., Jiang, Z., Qiu, Y., Zhang, S., Li, K., 2022. Thyroid nodule segmentation and classification in ultrasound images through intra-and inter-task consistent learning. Med. Image Anal. 79, 102443.

Kenton, J.D.M.W.C., Toutanova, L.K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S., 2023. One-shot weakly-supervised segmentation in 3D medical images. IEEE Trans. Med. Imaging.

Lester, B., Al-Rfou, R., Constant, N., 2021. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059.

Li, Y., Lao, Q., Kang, Q., Jiang, Z., Du, S., Zhang, S., Li, K., 2023a. Self-supervised anomaly detection, staging and segmentation for retinal images. Med. Image Anal. 87, 102805.

Li, Z.G., Li, B.C., Li, Z.W., Hu, H.Y., Ma, X., Cao, H., Yu, Z.H., Dai, H.P., Wang, J., Wang, C., et al., 2023b. The potential diagnostic biomarkers for the IgG subclass in coal workers' pneumoconiosis. J. Immunol. Res. 2023.

Li, J., Li, D., Savarese, S., Hoi, S., 2023c. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. PMLR, pp. 1–13.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Adv. Neural Inf. Process. Syst. 36.

Liu, H., Li, C., Wu, Q., Lee, Y.J., 2024. Visual instruction tuning. In: Advances in Neural Information Processing Systems, vol. 36.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Luo, G., Huang, M., Zhou, Y., Sun, X., Jiang, G., Wang, Z., Ji, R., 2023a. Towards efficient visual adaption via structural re-parameterization. arXiv preprint arXiv:2302.08106.

Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2022a. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal. 82, 102642.

Luo, X., Song, T., Wang, G., Chen, J., Chen, Y., Li, K., Metaxas, D.N., Zhang, S., 2022b. SCPM-net: An anchor-free 3D lung nodule detection network using sphere representation and center points matching. Med. Image Anal. 75, 102287.

Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R., 2023b. Cheap and quick: Efficient vision-language instruction tuning for large language models. arXiv preprint arXiv:2305.15023.

Ma, Y., Cui, W., Liu, J., Guo, Y., Chen, H., Li, Y., 2023. A multi-graph cross-attention based region-aware feature fusion network using multi-template for brain disorder diagnosis. IEEE Trans. Med. Imaging.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616 (7956), 259–265.

OpenAI, 2023a. ChatGPT. https://chat.openai.com. (Accessed 7 November 2023).

OpenAI, 2023b. GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al., 2020. Video-based AI for beat-to-beat assessment of cardiac function. Nature 580 (7802), 252–256.

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q., 2021. Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 367–376.

Qi, X.M., Luo, Y., Song, M.Y., Liu, Y., Shu, T., Liu, Y., Pang, J.L., Wang, J., Wang, C., 2021. Pneumoconiosis: current status and future prospects. Chin. Med. J. 134 (08), 898–907.

Qu, J., Wei, X., Qian, X., 2023. Generalized pancreatic cancer diagnosis via multiple instance learning and anatomically-guided shape normalization. Med. Image Anal. 86, 102774.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al., 2023. Large language models encode clinical knowledge. Nature 620 (7972), 172–180.

Stan, G.B.M., Rohekar, R.Y., Gurwicz, Y., Olson, M.L., Bhiwandiwalla, A., Aflalo, E., Wu, C., Duan, N., Tseng, S.Y., Lal, V., 2024. LVLM-intrepret: An interpretability tool for large vision-language models. arXiv preprint arXiv:2404.03118.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F., 2020. Learning to summarize with human feedback. Adv. Neural Inf. Process. Syst. 33, 3008–3021.

Sun, W., Wu, D., Luo, Y., Liu, L., Zhang, H., Wu, S., Zhang, Y., Wang, C., Zheng, H., Shen, J., et al., 2023. ExpertNet: Defeat noisy labels by deep expert consultation paradigm for pneumoconiosis staging on chest radiographs. Expert Syst. Appl. 120710.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023a. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al., 2023a. Sam-med3d. arXiv preprint arXiv:2310.15161.

Wang, L., Lin, Z.Q., Wong, A., 2020a. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Sci. Rep. 10 (1), 19549.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106.

Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., et al., 2023b. A real-world dataset and benchmark for foundation model adaptation in medical image classification. Sci. Data 1–9.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887.

Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S., 2023c. Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925.

Wang, X., Yu, J., Zhu, Q., Li, S., Zhao, Z., Yang, B., Pu, J., 2020b. Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. Occup. Environ. Med. 77 (9), 597–602.

Wen, S., Fang, G., Zhang, R., Gao, P., Dong, H., Metaxas, D., 2023. Improving compositional text-to-image generation with large vision-language models. arXiv preprint arXiv:2310.06311.

Wu, L., Gao, X., Hu, Z., Zhang, S., 2023a. Pattern-aware transformer: Hierarchical pattern propagation in sequential medical images. IEEE Trans. Med. Imaging.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023b. Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Xing, X., Chen, Z., Hou, Y., Yuan, Y., 2023. Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis. Med. Image Anal. 102874.

Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D., 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097.

Xu, X., Jia, Q., Yuan, H., Qiu, H., Dong, Y., Xie, W., Yao, Z., Zhang, J., Nie, Z., Li, X., et al., 2023a. A clinically applicable AI system for diagnosis of congenital heart diseases based on computed tomography images. Med. Image Anal. 90, 102953.

Xu, L., Ni, Z., Liu, X., Wang, X., Li, H., Zhang, S., 2023b. Learning a multi-task transformer via unified and customized instruction tuning for chest radiograph interpretation. arXiv preprint arXiv:2311.01092.

Yi, H., Qin, Z., Lao, Q., Xu, W., Jiang, Z., Wang, D., Zhang, S., Li, K., 2023. Towards general purpose medical AI: Continual learning medical foundation model. arXiv preprint arXiv:2303.06580.

You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B., 2023. CXR-CLIP: Toward large scale chest X-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 101–111.

Zhang, Y., Gao, J., Tan, Z., Zhou, L., Ding, K., Zhou, M., Zhang, S., Wang, D., 2024. Data-centric foundation models in computational healthcare: A survey. arXiv preprint arXiv:2401.02458.

Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D., 2023a. Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 272–282.

Zhang, R., Han, J., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Gao, P., Qiao, Y., 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199.

Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al., 2023c. Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514.

Zhang, S., Metaxas, D., 2024. On the challenges and perspectives of foundation models for medical image analysis. Med. Image Anal. 91, 102996.

Zheng, R., Deng, K., Jin, H., Liu, H., Zhang, L., 2019. An improved CNN-based pneumoconiosis diagnosis method on X-ray chest film. In: International Conference on Human Centered Computing. Springer, pp. 647–658.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022a. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 16816–16825.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022b. Learning to prompt for vision-language models. Int. J. Comput. Vis. 130 (9), 2337–2348.