LLaVA-RadZ: Can Multimodal Large Language Models Effectively Tackle Zero-shot Radiology Recognition?

Anonymous authors
Paper under double-blind review

Abstract

Recently, Multimodal Large Language Models (MLLMs) have demonstrated exceptional capabilities in visual understanding and reasoning across various vision-language tasks. However, we found that MLLMs cannot process effectively from fine-grained medical image data in the traditional Visual Question Answering (VQA) pipeline, as they do not exploit the captured features and available medical knowledge fully, results in MLLMs usually performing poorly in zero-shot medical disease recognition. Fortunately, this limitation does not indicate that MLLMs are fundamentally incapable of addressing fine-grained recognition tasks. From a feature representation perspective, MLLMs demonstrate considerable potential for tackling such challenging problems. Thus, to address this challenge, we propose ***LLaVA-RadZ***, a simple yet effective framework for zero-shot medical disease recognition via utilizing the existing MLLM features. Specifically, we design an end-to-end training strategy, termed *Decoding-Side Feature Alignment Training (**DFAT**)* to take advantage of the characteristics of the MLLM decoder architecture and incorporate modality-specific tokens tailored for different modalities. Additionally, we introduce a *Domain Knowledge Anchoring Module (**DKAM**)* to exploit the intrinsic medical knowledge of large models, which mitigates the *category semantic gap* in image-text alignment. Extensive experiments demonstrate that our LLaVA-RadZ significantly outperforms traditional MLLMs in zero-shot disease recognition, achieving the comparable performance to the well-established and highly-optimized CLIP-based approaches.

1 Introduction

With the rapid advancement of deep learning technologies, an increasing number of studies have focused on their applications in medical disease diagnosis, yielding remarkable results (Chan et al., 2020; Jamshidi et al., 2020; Lee et al., 2022; Tran et al., 2021). However, these approaches typically rely on high-quality annotations provided by clinical experts. Unlike natural image datasets, annotating medical images is both costly and time-consuming. To address this challenge, recent research has explored methods based on paired medical images and textual reports, leveraging contrastive learning techniques. By minimizing the distance between paired samples while maximizing the distance between unpaired ones, these CLIP-based approaches enable zero-shot disease recognition, thereby reducing reliance on extensive medical data annotation to a certain extent. In our in-depth investigation of advanced zero-shot disease recognition methods in the medical domain, several representative CLIP-based models (Lai et al., 2024; Wu et al., 2023; Zhang et al., 2023b; Phan et al., 2024) have achieved significant performance improvements leveraging the capabilities of Large Language Models (LLMs) or incorporate expert domain knowledge to some extent, rather than fully leveraging the models' intrinsic understanding capabilities.

Recently, Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Team et al., 2023; Liu et al., 2023; Huang et al., 2024; 2025; You et al., 2025) have demonstrated remarkable capabilities across various user-oriented vision-language tasks, such as image comprehension and reasoning, offering new possibilities for zero-shot disease recognition in medical applications. Among these, LLaVA-Med (Li et al., 2024a) has exhibited exceptional domain-specific medical knowledge in dialogue-based tasks, indicating that it possesses a certain degree of medical expertise. However,

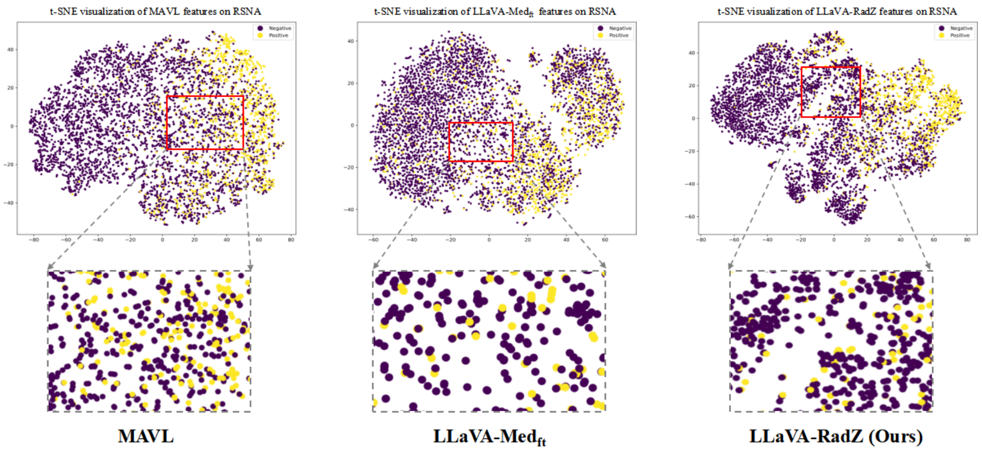


Figure 1: Comparison of Feature Distributions among MAVL, LLaVA-Med$_{ft}$, and LLaVA-RadZ on the RSNA Dataset.

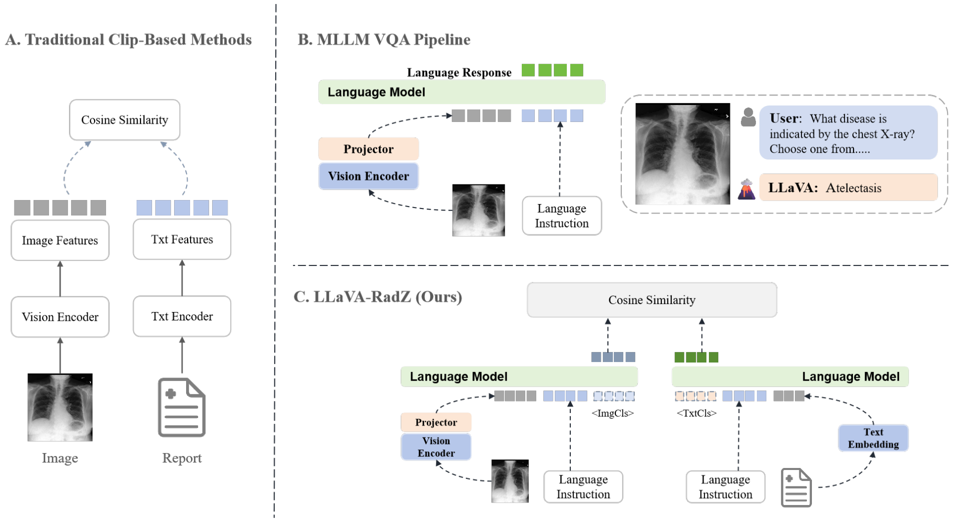


Figure 2: Framework comparison of traditional CLIP-based methods, MLLM VQA pipeline, and the proposed LLaVA-RadZ.

a recent study (Zhang et al., 2024) found that MLLMs, *i.e.*, LLaVA (Liu et al., 2023)) performed significantly worse than CLIP (Radford et al., 2021) on standard image classification tasks.

To further validate this observation, we conducted zero-shot classification experiments using multiple MLLMs on five medical imaging datasets (see Tab. 1). The experimental results are consistent with previous findings, confirming that MLLMs exhibit suboptimal performance in image classification, particularly when dealing with complex medical images. To enhance the generalization capability of MLLMs in radiology disease recognition tasks, we employed a fine-tuning strategy and performed supervised fine-tuning on the MIMIC-CXR dataset (Johnson et al., 2019). Additionally, inspired by the work of (Zhang et al., 2024), we incorporated a series of optimizations. While these improvements yielded performance gains, the results remained inferior compared to CLIP-based models. This phenomenon raises a critical question: *Can MLLMs effectively perform zero-shot disease recognition?*

As shown in Fig. 1, we visualize the feature distributions of MAVL (Phan et al., 2024), LLaVA-Med$_{ft}$ (fine-tuned by the same dataset of our LLaVA-RadZ) and LLaVA-RadZ on the RSNA (Shih et al., 2019) dataset. The results indicate that MLLM exhibits strong feature extraction capabilities, comparable to the well-established MAVL in the domain. However, in the disease recognition task, MAVL significantly outperforms fine-tuned LLaVA-Med. We hypothesize that this performance gap arises because MLLMs fail to fully utilize the extracted features for effective disease identification via traditional VQA pipeline.

Inspired by this, we propose a simple yet effective LLaVA-RadZ framework for zero-shot disease recognition using the MLLM features. Our proposed framework has the fundamental difference compared with previous CLIP-base methods and traditional MLLM VQA pipeline. As shown in Fig. 2, we design a dedicated MLLM feature-based framework to address zero-shot medical disease recognition. Our proposed framework effectively leverages pre-trained MLLM representations to overcome the inherent limitations of the traditional VQA pipeline on this task. Specifically, firstly, we introduce a new training strategy, Decoding-Side Feature Alignment Training (DFAT). Specifically, we introduce special tokens for both image and text modalities and leverage the autoregressive generation capability of the decoder architecture to extract global representations of images and texts. Additionally, we incorporate a cross-modal contrastive loss to optimize the model's ability to learn discriminative features. Furthermore, to mitigate the semantic category gap encountered during fine-grained alignment between medical images and textual reports, we design a Domain Knowledge Anchoring Module (DKAM). DKAM utilizes the model's intrinsic medical knowledge to extract the semantic information underlying disease categories, constructing disease description vectors that serve as an intermediary bridge to facilitate the alignment between medical images and textual reports, thereby establishing a stable relationship. To further enhance the correlation among medical images, textual reports, and disease categories, a category knowledge-guided loss strengthens the association between similar images and corresponding textual reports.

Our main contributions can be summarized as follows.

- We analyze the limitations of current MLLMs in addressing complex fine-grained medical disease recognition tasks, investigate the underlying causes of these constraints, and propose a novel end-to-end feature-based MLLM framework to mitigate these challenges. To the best of our knowledge, we are the ***first*** work in the field of medical disease recognition to explore how to use MLLM features directly to solve complex recognition problems.
- We propose the tailored training strategy DFAT, and incorporate a cross-modal contrastive loss to optimize the model's ability to achieve effective alignment between visual and textual features. Furthermore, we design a DKAM to leverage MLLM's intrinsic medical knowledge and effectively mitigate semantic gap in image-text alignment, thereby enhancing category-level alignment.
- We conduct extensive experiments on multiple large-scale radiology diagnosis datasets, validating the potential of LLaVA-RadZ in zero-shot disease recognition tasks.

2 Approach

2.1 Can Med-LLMs Be Good Medical Classifiers?

Previous studies have explored the classification capabilities of multimodal large language models (MLLMs), revealing that their performance on image classification tasks is often limited. For example, (Zhang et al., 2024) investigates the performance differences in classification between MLLMs and CLIP, focusing on factors such as inference strategies, training approaches, and datasets. Inspired by this work, we extend the exploration to zero-shot tasks in the medical domain. Unlike natural images and text, the relationship between medical images and reports is more complex. We seek to investigate whether large medical models, leveraging domain-specific knowledge, can achieve superior performance on medical zero-shot tasks.

We first evaluated two open-source MLLMs, i.e., LLaVA-1.5 (Liu et al., 2023) and LLaVA-Med (Li et al., 2024a), on five medical datasets in a zero-shot classification setting. The evaluation followed a general large-model classification approach, where the model selects the correct category from a set of candidate options. As shown in Tab. 1, these models demonstrated limited performance in disease classification tasks and failed to accurately identify various medical conditions. Given the potential knowledge limitations of these models, we further assessed the performance of more powerful proprietary MLLMs (i.e., Qwen2.5-Max (Yang et al., 2024), Gemini-Pro (Team et al., 2023), and GPT-4o (Achiam et al., 2023)) on zero-shot medical disease recognition tasks. As shown in table 1, these models exhibited superior classification capabilities. However, they still lagged behind the state-of-the-art domain-specific methods in medical classification.

To enhance the generalization ability of MLLMs in radiology disease identification, we conducted Supervised Fine-Tuning (SFT) on LLaVA-1.5 (Liu et al., 2023) and LLaVA-Med (Li et al., 2024a) using the publicly available MIMIC-CXR dataset (Johnson et al., 2019). Surprisingly, the fine-tuned models did not achieve consistent performance improvements across the five datasets. In some cases, their classification performance even deteriorated. Further analysis of the model outputs revealed that MLLMs did not always focus on disease-specific information in radiology reports. Instead, they tended to overlearn the textual structures and linguistic patterns of the reports, which limited their classification capability. To mitigate this issue, we incorporated the Chain-of-Thought (CoT) prompting strategy and adjusted the model's reasoning approach, inspired by the methodology of (Zhang et al., 2024), to optimize the model's decision-making process. This approach led to moderate improvements in classification performance on medical datasets. Although the models have not yet reached optimal performance, the results suggest that MLLMs still hold significant potential for zero-shot medical disease recognition.

2.2 Motivation

As previously discussed, despite possessing a certain level of domain knowledge, medical MLLMs have not yet demonstrated remarkable performance in zero-shot medical tasks. Even with further instruction tuning, their performance remains inferior to that of existing vision-language models (VLMs). However, it is noteworthy that modifying the inference strategy leads to significant performance improvements, suggesting that MLLMs are indeed capable of capturing medical image and text features. Nevertheless, these features have yet to be fully exploited.

To address this limitation, we propose the LLaVA-RadZ framework, introducing a novel end-to-end training strategy, Decoding-Side Feature Alignment Training (DFAT). This approach leverages the unique properties of the MLLM decoder architecture while incorporating modality-specific special tokens to facilitate effective interaction between medical images and textual features, ultimately achieving more robust cross-modal alignment. As illustrated in Fig. 1, we compare the feature distribution of our model with MAVL (Phan et al., 2024), the current state-of-the-art method, on the RSNA (Shih et al., 2019) dataset. The results clearly demonstrate that our model achieves better clustering of intra-class samples while enhancing inter-class separation, validating the effectiveness of our approach. Furthermore, we introduce the Domain Knowledge Anchor Module (DKAM), which harnesses the intrinsic medical knowledge of LLMs to bridge the semantic gap between images and text, enabling more precise disease classification.

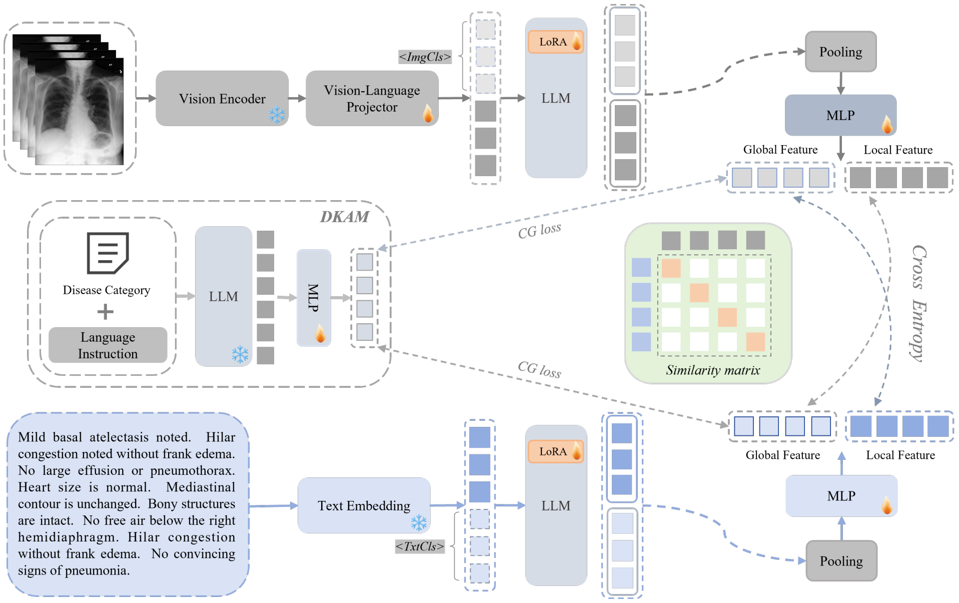
2.3 The Proposed LLaVA-RadZ

We aim to learn generalizable medical image representations from radiology reports to enhance various downstream medical image recognition tasks, particularly when labeled data is scarce. The overall framework is illustrated in Fig. 3. Given a pair of medical images and reports, the image and text are first passed through separate visual and text encoders to obtain their respective encoded features. These encoded features and specially designed tokens are then fed into a language model to obtain the final feature representation. The features are mapped into a common representational space via an MLP projection layer and optimized with the InfoNCE loss. Furthermore, we propose a Domain Knowledge Anchor Module (DKAM), which leverages domain knowledge inherent in the model to guide the alignment of text and image features at the category level.

2.3.1 End-to-End Training Strategy

Currently, most MLLMs employ generation-based training objectives for instruction fine-tuning. Although this approach effectively captures the features of medical images and textual reports, its performance in zero-shot tasks remains suboptimal, as it fails to fully leverage these features. To address this issue, we propose a novel training strategy, Decoding-Side Feature Alignment Training (DFAT), as illustrated in Fig. 3.

We consider a training dataset consisting of N pairs of medical image-text samples, denoted as $S_{\text{train}} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$. The medical image $X_i \in \mathbb{R}^{H \times W \times 3}$, with H and W repre-



Mild basal atelectasis noted. Hilar congestion noted without frank edema. No large effusion or pneumothorax. Heart size is normal. Mediastinal contour is unchanged. Bony structures are intact. No free air below the right hemidiaphragm. Hilar congestion without frank edema. No convincing signs of pneumonia.

Figure 3: The LLaVA-RadZ framework consists of three components. (A) Construct a category semantic vector repository using the domain knowledge anchoring module (DKAM). (B) Encode medical images and text, appending $\langle ImgCls \rangle$ and $\langle TxtCls \rangle$ tokens before feeding them into the LLM. (C) Extract global and local features, optimizing with cross-entropy loss, while leveraging the semantic repository for category-level alignment.

senting the height and width of the image, respectively. Y_i refers to the corresponding medical text report associated with the image.

Specifically, we design special tokens for both image and text modalities, where $< ImgCls_i > (i = 0, .., 4)$ denotes image feature tokens and $< TxtCls_i > (i = 0, .., 8)$ denotes text feature tokens. These special tokens are attached to the image prompt X_{prompt} and the text prompt Y_{prompt}, respectively. The image prompt X_{prompt} has a format similar to “What disease is indicated by the chest X-ray?”, while the text prompt Y_{prompt} follows a format such as “What disease is described in this text?”. By appending special tokens, we obtain the modified prompts $\tilde{X}_{\text{prompt}}$ and $\tilde{Y}_{\text{prompt}}$, which is represented as:

$$\tilde{X}_{\text{prompt}} = X_{\text{prompt}} + < ImgCls_i >_{(i=0,..,4)}, \tag{1}$$

$$\tilde{Y}_{\text{prompt}} = Y_{\text{prompt}} + < TxtCls_i >_{(i=0,..,8)}. \tag{2}$$

When an image and its corresponding prompt $\tilde{X}_{\text{prompt}}$ are input into the MLLM $\mathcal{F}$ to generate a response $\hat{R}_{\text{img}}$. Similarly, when a text sample and its corresponding feature extraction prompt $\tilde{Y}_{\text{prompt}}$ are provided as input, the model produces a response $\hat{R}_{\text{txt}}$. This process can be formally expressed as:

$$\hat{R}^i_{\text{img}} = \mathcal{F}(X_i, \tilde{X}_{\text{prompt}}), \quad \hat{R}^i_{\text{txt}} = \mathcal{F}(Y_i, \tilde{Y}_{\text{prompt}}). \tag{3}$$

Due to the autoregressive nature of the decoder architecture, when the LLM processes visual and textual information to generate responses, its internal representations are stored in the designated special tokens. Specifically, we extract the penultimate layer embedding $\tilde{h}_{\text{img}}$ corresponding to the special token $< ImgCls_i >$, which stores the global image features $H^{\text{global}}_{\text{img}} \in \mathbb{R}^{B \times I \times K}$. Here, B denotes the number of image-text pairs in each batch, I represents the number of special image tokens, and K is the dimension of the shared embedding space. After applying a pooling operation followed by an MLP projection layer γ_{img}, we obtain the global image feature representation $X_g \in \mathbb{R}^{B \times K}$. The local image feature $X_l \in \mathbb{R}^{B \times K}$ is obtained by pooling the hidden states of all tokens except those corresponding to special tokens, followed by an MLP projection layer γ_{img}:

$$X_g = \gamma_{\text{img}}(\text{AvgPool}(H^{\text{global}}_{\text{img}})), \quad X_l = \gamma_{\text{img}}(\text{AvgPool}(H^{\text{local}}_{\text{img}})). \tag{4}$$

Similarly, we extract the global text representation $Y_g \in \mathbb{R}^{B\times K}$ and the local text representation $Y_l \in \mathbb{R}^{B\times K}$ using the same methodology:

$$Y_g = \gamma_{\text{txt}}(\text{AvgPool}(H_{\text{txt}}^{\text{global}})), \quad Y_l = \gamma_{\text{txt}}(\text{AvgPool}(H_{\text{txt}}^{\text{local}})). \tag{5}$$

To further enhance fine-grained alignment across different modalities, we introduce a cross-modal contrastive loss, L_{CA}. Specifically, for the i-th image-text pair $(X_i,\, Y_i)$ in a batch, we alternately align the global and local features of images and texts. This procedure yields two symmetric, temperature-normalized InfoNCE objectives: one aligns global image features with local text features, and the other aligns local image features with global text features. These objectives maximize the mutual information between image-text pairs in the latent space.

For the alignment between global image features and local text features, we calculate two similarity matrices, $S_i^{X_g\to Y_l}$ and $S_i^{Y_l\to X_g}$, with the following computation:

$$S_i^{X_g\to Y_l} = \frac{X_{g,i}\cdot Y_{l,i}^T}{\tau}, \quad S_i^{Y_l\to X_g} = \frac{Y_{l,i}\cdot X_{g,i}^T}{\tau}. \tag{6}$$

where τ is the temperature hyperparameter. Subsequently, we compute the contrastive loss between the global image and the local text, with the following formula:

$$L_{\text{CA}}^{X_g\to Y_l,i} = -\log\frac{\exp\big(S_i^{X_g\to Y_l}\big)}{\sum_{k=1}^{B}\exp\big(S_k^{X_g\to Y_l}\big)}, \quad L_{\text{CA}}^{Y_l\to X_g,i} = -\log\frac{\exp\big(S_i^{Y_l\to X_g}\big)}{\sum_{k=1}^{B}\exp\big(S_k^{Y_l\to X_g}\big)}. \tag{7}$$

$$L_{\text{CA}}^{X_g\to Y_l} = \frac{1}{2}\sum_{i=1}^{B}\Big(L_{\text{CA}}^{X_g\to Y_l,i} + L_{\text{CA}}^{Y_l\to X_g,i}\Big). \tag{8}$$

Similarly, for the alignment between local image features and global text features, we compute the contrastive loss between the local image and global text.

$$L_{\text{CA}}^{X_l\to Y_g} = -\frac{1}{2}\sum_{i=1}^{B}\Big(\log\frac{\exp\big(S_i^{X_l\to Y_g}\big)}{\sum_{k=1}^{B}\exp\big(S_k^{X_l\to Y_g}\big)} + \log\frac{\exp\big(S_i^{Y_g\to X_l}\big)}{\sum_{k=1}^{B}\exp\big(S_k^{Y_g\to X_l}\big)}\Big). \tag{9}$$

Finally, we obtain our cross-modal contrastive loss L_{CA}.

$$L_{CA} = \frac{1}{2}\left(L_{\text{CA}}^{X_g\to Y_l} + L_{\text{CA}}^{X_l\to Y_g}\right). \tag{10}$$

2.3.2 Domain Knowledge Anchor Module

In aligning medical images with text reports, we observed that the critical entity of the medical disease categories was merely encoded as features by the model, without considering the underlying semantics. To address this limitation and further enhance fine-grained alignment capabilities, we introduce the Domain Knowledge Anchoring Module (DKAM). Initially, we leverage the inherent medical domain expertise of an LLM to generate descriptive explanations for each disease category. These generated disease descriptions serve as an intermediary bridge to guide the alignment between medical images and text reports. Specifically, we input the disease list D_{list} from the training dataset along with a designed prompt template K_{prompt} into the LLM $\mathcal{F}$. This process is formally expressed as:

$$\hat{R}_{\text{dis}} = \mathcal{F}(D_{\text{list}}, K_{\text{prompt}}). \tag{11}$$

By fully harnessing the LLM's exceptional semantic understanding, we prompt the model to explore the underlying semantics of the disease categories and discern their distinctions, ultimately producing a refined disease description. The features extracted from the LLM's response are then mapped via a multi-layer perceptron (MLP) to yield the disease description vector $\hat{D}$, which is represented as:

$$\hat{D} = \gamma_{\text{dis}}\left(\hat{R}_{\text{dis}}\right). \tag{12}$$

Subsequently, we introduce the Category of Knowledge-guided Contrastive Loss L_{CG}. Specifically, we calculate the cross-entropy loss between the disease description vector $\hat{D}$ and the global features of both the images X_g and the text Y_g. This design encourages the model to better capture the

Table 1: Comparison of zero-shot disease classification performance of public MLLMs and LLaVA-based exploratory methods across five medical benchmarks. “ft” denotes supervised fine-tuning with LoRA, “CoT” refers to zero-shot chain-of-thought prompting templates, and “Inference” represents CLIP inference strategies. The best results are highlighted in bold and the second-best results are underlined.

Dataset		CheXpert			ChestXray-14			COVIDx CXR-2			RSNA Pneumonia			SIIM-ACR		
Method	**Model**	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑
MLLM	LLaVA-1.5 (7B) (Liu et al., 2023)	-	7.50	8.28	-	3.33	6.92	-	53.14	50.28	-	40.53	55.34	-	23.66	50.36
	LLavA-Med (7B) (Li et al., 2024a)	-	6.87	8.94	-	8.02	6.78	-	34.90	50.03	-	18.58	50.00	-	21.91	49.90
	Qwen2.5-Max (Yang et al., 2024)	-	32.23	67.97	-	19.04	76.19	-	75.91	76.81	-	43.58	43.59	-	64.70	72.57
	Gemini-Pro (Team et al., 2023)	-	35.01	76.08	-	14.16	77.78	-	62.84	62.90	-	44.23	51.43	-	61.43	72.03
	GPT-4o (Achiam et al., 2023)	-	45.85	81.14	-	19.85	81.55	-	50.93	**77.08**	-	54.20	65.33	-	64.57	72.11
Explorative Methods	LLaVA-1.5-7B$_{\mathbf{ft}}$	-	10.61	19.62	-	7.85	19.06	-	27.74	25.18	-	43.60	34.80	-	52.37	50.95
	LLavA-Med-7B$_{\mathbf{ft}}$	-	14.25	31.46	-	9.00	21.43	-	27.42	24.09	-	46.72	38.88	-	53.11	57.68
	LLavA-Med-7B$_{\mathbf{ft}}$ + CoT (Zhang et al., 2024)	-	8.90	26.23	-	8.33	20.46	-	27.12	26.55	-	49.59	43.80	-	54.06	51.07
	LLavA-Med-7B$_{\mathbf{ft}}$ + Inference (Zhang et al., 2024)	71.00	44.85	75.45	64.30	21.73	70.86	71.07	69.84	60.39	77.51	69.85	72.90	71.25	68.26	71.27
Ours	**LLaVA-RadZ**$_{\mathbf{ft}}$	**73.36**	**48.59**	**82.15**	**72.61**	**27.91**	**84.64**	**84.36**	**77.53**	74.58	**86.98**	**76.18**	**83.28**	**89.92**	**79.57**	**84.38**

Table 2: Comparison of performance with other SOTA methods on four medical datasets for the zero-shot classification task, with AUC, F1, and ACC scores reported. The best results are highlighted in bold and the second-best results are underlined.

Method	ChestXray-14			COVIDx CXR-2			RSNA Pneumonia			SIIM-ACR		
	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑
ConVIRT (Zhang et al., 2022)	53.15	12.38	57.88	62.78	71.23	63.84	79.21	55.67	75.08	64.25	42.87	53.42
GLoRIA (Huang et al., 2021)	55.92	14.20	59.47	64.52	70.78	60.21	70.37	48.19	70.54	54.71	40.39	47.15
BioViL (Boecking et al., 2022)	57.82	15.64	61.33	61.40	70.92	58.20	84.12	54.59	74.43	70.28	46.45	68.22
CheXzero (Tiu et al., 2022)	66.99	21.99	65.38	73.13	76.13	71.45	85.13	61.49	78.34	84.60	65.97	77.34
MedKLIP (Wu et al., 2023)	72.33	24.18	79.40	76.28	76.54	71.96	86.57	63.28	79.97	89.79	72.73	83.99
MAVL (Phan et al., 2024)	**73.50**	26.25	82.77	83.86	**81.73**	**78.07**	86.91	63.41	82.42	**92.04**	77.95	**87.14**
Ours	72.61	**27.91**	**84.64**	**84.36**	77.53	74.58	**86.98**	**76.18**	**83.28**	89.92	**79.57**	84.38

semantic relationships among images, text, and disease categories during training, achieving a more robust category-level alignment.

$$S_i^{\text{img-disease}} = \frac{X_{g,i} \cdot D^T}{\tau}, \quad S_i^{\text{txt-disease}} = \frac{Y_{g,i} \cdot D^T}{\tau}. \tag{13}$$

$$L_{\text{CG},i}^{\text{txt}} = -\log \frac{\exp\left(S_i^{\text{txt-disease}}\right)}{\sum_{k=1}^{N} \exp\left(S_k^{\text{txt-disease}}\right)}, \quad L_{\text{CG},i}^{\text{img}} = -\log \frac{\exp\left(S_i^{\text{img-disease}}\right)}{\sum_{k=1}^{N} \exp\left(S_k^{\text{img-disease}}\right)}. \tag{14}$$

Here, N represents the number of disease categories, B denotes the number of medical image-text pairs in each batch, and τ is the temperature hyperparameter. The final category of knowledge-guided loss is as follows:

$$L_{\text{CG}} = \frac{1}{2} \sum_{i=1}^{B} \left(L_{\text{CG},i}^{\text{txt}} + L_{\text{CG},i}^{\text{img}} \right). \tag{15}$$

By combining the category knowledge-guided loss and the cross-modal contrastive loss, the final objective function is defined as follows:

$$L_{\text{total}} = \lambda L_{\text{CA}} + (1 - \lambda) L_{\text{CG}}, \tag{16}$$

where λ is a balancing factor used to adjust the weights of the two losses, and it is set to 0.5 by default.

3 Experiments

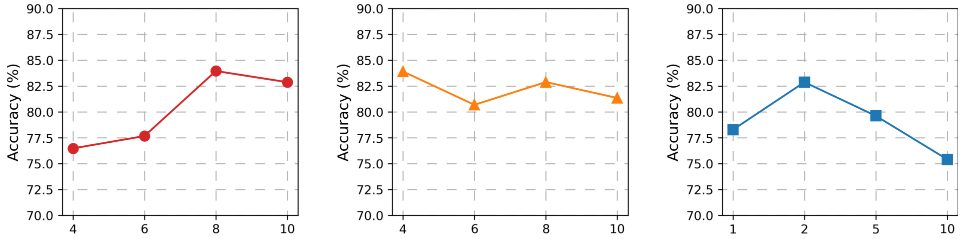
In this section, we first provide an overview of the dataset employed in our experiments, including those used for pre-training and the various downstream tasks. Subsequently, we outline the implementation details and describe the baselines considered for comparison.

3.1 Dataset

In our experiments, we pre-trained the model using the MIMIC-CXR dataset (Johnson et al., 2019). For downstream tasks, we primarily evaluated the model’s performance in medical disease classification using multiple benchmark datasets, including ChestX-ray14 (Wang et al., 2017), RSNA Pneumonia (Shih et al., 2019), SIIM-ACR Pneumothorax (sii, 2019), CheXpert (Irvin et al., 2019), and COVIDx CXR-2 (Pavlova et al., 2022). Detailed information on these datasets can be found in the supplementary material.

Table 3: Comparison of performance with other SOTA methods at different data portions for fine-tuning classification task. AUC scores are reported. The best results are highlighted in bold and the second-best results are underlined.

Method	**RSNA Pneumonia**			**Pneumothorax**			**COVIDx CXR-2**		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	68.94	83.31	87.12	53.11	76.18	87.48	85.11	93.65	98.86
ConVIRT (Zhang et al., 2022)	78.86	85.42	87.64	72.39	80.41	91.67	90.30	97.74	99.70
GLoRIA (Huang et al., 2021)	79.13	85.59	87.83	75.85	86.20	91.89	92.74	97.18	99.54
BioViL (Boecking et al., 2022)	80.27	86.04	88.29	70.29	79.45	88.05	92.39	98.39	99.68
MedKLIP (Wu et al., 2023)	82.11	87.14	88.58	85.24	89.91	93.02	95.58	98.77	99.77
MAVL (Phan et al., 2024)	86.09	87.90	88.94	**91.53**	**93.00**	94.48	97.18	99.15	99.90
Ours	**88.23**	**88.57**	**89.49**	88.42	89.96	**94.50**	**98.32**	**99.80**	**99.96**



(a) Number of Text Special Tokens (b) Number of Image Special Tokens (c) Depth of Hidden Layers

Figure 4: Effect of Special Token Numbers and Hidden Layer Depth on ChestXray-14 Classification.

3.2 Evaluation Metrics

For the zero-shot classification task, we employ standard classification evaluation metrics, including Accuracy, AUC score, and F1 score. The macro-average metrics are reported for all diseases present in the target dataset.

3.3 Zero-shot evaluation

As shown in Tab. 2, we compare the performance of established methods in the field on the zero-shot classification task for radiological diseases, evaluated on four officially released test datasets. Our findings demonstrate that, compared to conventional CLIP-style models such as ConVIRT (Zhang et al., 2022), GLoRIA (Huang et al., 2021), BioViL (Boecking et al., 2022), and CheXzero (Tiu et al., 2022), our approach exhibits significant advantages. Even when compared to state-of-the-art models incorporating external models or domain-specific expert knowledge, our method remains highly competitive. Specifically, on the multi-class dataset ChestXray-14, our model surpasses the supervised learning method MAVL (Phan et al., 2024) by 1.87% in accuracy. Moreover, on the RSNA Pneumonia dataset, we achieve a 12.77% improvement in F1 score. These results indicate that multimodal large language models (MLLMs) possess strong feature extraction capabilities, further underscoring their immense potential in medical disease classification tasks.

3.4 Fine-tuning evaluation

Consistent with previous studies (Phan et al., 2024; Wu et al., 2023), we fine-tune the model on downstream datasets using 1%, 10%, and 100% of the available data and further evaluate its performance. Tab. 3 presents the fine-tuning results across three datasets, demonstrating that our model consistently maintains a competitive advantage. Notably, when fine-tuned with only 1% data, our proposed LLaVA-RadZ outperforms the MAVL (Phan et al., 2024) model by 2.14% on the RSNA Pneumonia and by 1.14% on COVIDx. Even when fine-tuned with 100% data, our model continues to deliver performance improvements. This enhancement is likely attributed to our decoder-side alignment training strategy, which effectively captures global modality information and leverages the interaction between global and local features to achieve fine-grained cross-modal alignment, further strengthening the model's disease recognition capability.

Table 4: Ablation study of DKAM on ChestXray-14. D_1 represents a semantic vector library of 75 medical entities, and D_2 represents a semantic vector library of 14 disease categories.

#	*DKAM*	D_1	D_2	AUC ↑	F1 ↑	ACC ↑
a				69.31	27.30	82.32
b	✓	✓		68.67	25.73	81.84
c	✓		✓	**72.61**	**27.91**	**84.64**

Table 5: Ablation study of feature representations on ChestXray-14.

#	*Global*	*Local*	*Prompt*	AUC ↑	F1↑	ACC↑
a		✓		67.14	25.11	77.82
b		✓	✓	68.29	26.42	78.63
c	✓		✓	70.13	26.22	82.50
d	✓	✓	✓	**72.61**	**27.91**	**84.64**

3.5 Ablation Study

Ablation Study of DKAM. To validate the effectiveness of our proposed Domain Knowledge Anchor Module (DKAM), we conducted an ablation study on the ChestXray-14 dataset. With DKAM incorporated, we further investigated the impact of different category semantic vector repositories on the model's fine-grained alignment capability. Consistent with the previous MedKLIP study, we selected 75 primary medical entities from the MIMIC-CXR dataset. However, unlike MedKLIP, we leveraged the model's intrinsic domain knowledge to construct a category semantic vector repository, denoted as D_1. Additionally, we built a disease-specific semantic vector repository for the 14 medical disease categories present in the MIMIC-CXR training dataset, denoted as D_2.

As shown in Tab. 4 (a vs. c), the introduction of DKAM significantly enhances model performance. Using disease category semantics as an intermediary facilitates more precise alignment between medical images and textual descriptions at the category level. Further comparisons in Tab. 4 (b vs. c) demonstrate that, compared to a larger repository of medical entities, a semantic vector repository focusing on primary disease categories provides stronger guidance for image-text alignment. Moreover, additional medical entities in D_1, such as tip, tube, PICC, and device, may introduce noise and negatively impact alignment at the disease category level. This adverse effect is corroborated by the comparative results in Tab. 4 (a vs. b).

Ablation Study of Special Tokens. As shown in Tab. 1, we have demonstrated the effectiveness of the Decoding-Side Feature Alignment Training (DFAT) strategy. To further investigate the design of the critical special tokens integral to this approach, we conducted an in-depth analysis on the ChestXray-14 dataset. As illustrated in Fig. 4, we observed that the number of text and image tokens significantly influences model performance, with both an excessive and an insufficient count potentially resulting in a loss of modal information. Moreover, our study indicates that the optimal global features are not stored in the final hidden layer but rather in the penultimate layer, which may be attributed to the loss of fine-grained information due to deeper feature aggregation, thereby affecting overall performance.

Ablation Study of Features. During the process of cross-modal alignment, we conducted a detailed analysis of the impact of global and local features on model performance, and further investigated the effectiveness of using prompts, as shown in Tab. 5. The experimental results indicate that utilizing only local features yields the poorest performance, while relying solely on global features provides a certain advantage over local features. This may be attributed to the fact that the specially designed tokens for each modality can more precisely capture the global information of the corresponding modality. Moreover, the combination of global and local features achieves the best performance. Additionally, the incorporation of prompts further enhances the model's ability to capture feature information.

4 Conclusion

This paper proposes a simple yet effective framework, LLaVA-RadZ, for zero-shot medical disease recognition. First, we introduce an end-to-end decoding-side feature alignment training strategy to leverage the characteristics of the MLLM architecture and effectively store modality-related information. Additionally, we employ cross-modal contrastive learning to optimize feature alignment across modalities, enhancing the model's cross-modal understanding capabilities. Furthermore, we propose a domain knowledge anchoring Module to facilitate category-level alignment between medical images and textual descriptions. Experimental results demonstrate that LLaVA-RadZ achieves outstanding performance across multiple benchmarks, highlighting the significant potential of MLLMs in tackling zero-shot radiological disease recognition tasks.

Ethics Statement

This study follows the ICLR Code of Ethics. All experiments and data usage comply with relevant laws, regulations, and ethical requirements. The data used are from publicly available datasets or obtained with proper authorization, and have been appropriately preprocessed to ensure privacy and security. This work aims to advance scientific research and is not intended for any harmful or inappropriate applications. The authors declare no conflict of interest.

Reproducibility Statement

We provide detailed descriptions of our methods, datasets, and experimental settings in the main text and appendix. All source code and data processing scripts will be made publicly available upon publication to facilitate reproducibility. Additional implementation details are available in the supplementary materials.

References

Society for imaging informatics in medicine: Siim-acr pneumothorax segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation, 2019.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Piroon Jenjaroenpun, Thomas Powell, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri Blake, et al. Chest imaging representing a covid-19 positive rural us population. *Scientific data*, 7(1):414, 2020.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*, 2025.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaoshen Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. *arXiv preprint arXiv:2412.00876*, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Mohammad Jamshidi, Ali Lalbakhsh, Jakub Talla, Zdeněk Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi, Luigi La Spada, Mirhamed Mirmozafari, Mojgan Dehghani, et al. Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. *Ieee Access*, 8:109581–109595, 2020.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11137–11146, 2024.

Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R Rogers, Wendy K Chung, and Chunhua Weng. Deep learning for rare disease: A scoping review. *Journal of biomedical informatics*, 135:104227, 2022.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. *arXiv preprint arXiv:2402.18344*, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15 (6), 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yanming Liu, Xinyue Peng, Tianyu Du, Jianwei Yin, Weihao Liu, and Xuhong Zhang. Era-cot: improving chain-of-thought through entity relationship analysis. *arXiv preprint arXiv:2403.06932*, 2024.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pp. 304–323. Springer, 2024.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18798–18806, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Maya Pavlova, Naomi Terhljan, Audrey G Chung, Andy Zhao, Siddharth Surana, Hossein Aboutalebi, Hayden Gunraj, Ali Sabri, Amer Alaref, and Alexander Wong. Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Frontiers in Medicine*, 9: 861680, 2022.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome medicine*, 13:1–17, 2021.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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, 2017.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21372–21383, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ling You, Wenxuan Huang, Xinni Xie, Xiangyi Wei, Bangyan Li, Shaohui Lin, Yang Li, and Changbo Wang. Timesoccer: An end-to-end multimodal large language model for soccer commentary generation. *arXiv preprint arXiv:2504.17365*, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023b.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A RELATED WORK

Multi-modal Large Language Models. Inspired by the exceptional reasoning capabilities of large language models (LLMs), researchers are actively exploring ways to extend these abilities to the visual domain, driving advancements in multimodal LLMs. With the release of GPT-4 (Vision) (Achiam et al., 2023) and Gemini (Team et al., 2023), these models have demonstrated remarkable multimodal understanding and generation capabilities, further fueling research in this field.

To bridge the gap between vision encoders and LLMs, BLIP-2 (Li et al., 2023a) introduces a Q-Former that transforms image features into a format compatible with LLMs, enabling seamless integration with text embeddings. LLaVA (Liu et al., 2023) and MiniGPT-4 (Zhu et al., 2023) further enhance generalization and task performance by leveraging large-scale multimodal pretraining, followed by instruction tuning for specific applications. In the medical domain, LLMs have shown immense potential for advancing research and practical applications. Med-Flamingo (Moor et al., 2023) extends Flamingo to the medical field by pretraining on multimodal knowledge sources spanning multiple medical disciplines. LLaVA-Med (Li et al., 2024a) refines image-text pairs from PMC-15M (Zhang et al., 2023a) and trains a biomedical-specialized MLLM using a limited dataset, building upon the pre-trained parameters of LLaVA. Similarly, Med-PaLM (Singhal et al., 2023) fine-tunes PaLM (Chowdhery et al., 2023) using domain-specific medical instructions, demonstrating strong performance under human evaluation frameworks. Other notable models, such as Chat-Doctor (Li et al., 2023b) and Med-Alpaca (Han et al., 2023), have been tailored for medical question-answering and dialogue applications.

Despite the significant progress of MLLMs, several challenges remain (McKinzie et al., 2024; Tong et al., 2024; Zhang et al., 2024; He et al., 2025). Recent studies (Zhang et al., 2024; He et al., 2025) highlight the suboptimal performance of MLLMs in image classification, particularly in fine-grained category recognition. We find that this issue is especially pronounced in the medical domain, where precise classification is crucial for medical applications. To address these shortcomings, we are refining traditional MLLM training paradigms to enhance classification performance and improve fine-grained category comprehension.

Prompt Engineering. Prompting enhances the ability of pre-trained large language models (LLMs) to understand tasks by incorporating language instructions into the input text (Mondal et al., 2024; Shao et al., 2024; Liu et al., 2024; Li et al., 2024b). Recently, prompt-based techniques have also been applied to vision-language models to improve performance. In medical vision-language models (VLMs), GloRIA (Huang et al., 2021) generates a set of textual prompts to describe potential subtypes, severity levels, and anatomical locations for each disease category. MedKLIP (Wu et al., 2023) enhances model performance by retrieving descriptions of medical entities from the UMLS knowledge base (Bodenreider, 2004). CARZero (Lai et al., 2024) introduces a prompt-alignment strategy based on LLMs, integrating prompt templates into the training dataset to ensure alignment during both training and inference. MAVL (Phan et al., 2024) uses visual descriptions of pathological features to guide the model in effectively detecting diseases in medical images.

Although these approaches have successfully improved model performance through prompt-based strategies, they all rely on external models or expert knowledge, without fully leveraging the model's intrinsic understanding capabilities. Fortunately, recent research on LLaVA-Med (Li et al., 2024a) has demonstrated remarkable domain-specific conversational abilities, proving that it possesses a certain level of medical knowledge. Building upon LLaVA-Med (Li et al., 2024a), we further explore the feasibility of utilizing the model's inherent comprehension to enhance zero-shot medical classification performance.

B DATASET DETAILS

MIMIC-CXR v2 (Johnson et al., 2019). In our experiments, we pre-trained the model using the MIMIC-CXR, a publicly available collection of chest radiographs paired with corresponding radiology text reports. The MIMIC-CXR dataset comprises $377,110$ images corresponding to 227,835 radiographic studies from 65,379 patients. Since all downstream tasks utilize frontal-view images, we exclude all lateral-view images from the dataset. Moreover, we selectively retain only the findings and impressions sections from these reports.

ChestX-ray14 (Wang et al., 2017). ChestX-ray14 consists of $112,120$ frontal-view chest X-ray images from 30,805 unique patients, collected between 1992 and 2015. The official test set, comprising $22,433$ images, has been meticulously annotated by board-certified radiologists. For evaluation purposes, we restrict our testing to the official test set.

RSNA Pneumonia (Shih et al., 2019). RSNA Pneumonia includes over $260,000$ frontal-view chest X-rays with annotated pneumonia masks, collected by the Radiological Society of North America (RSNA). This dataset supports both pneumonia segmentation and classification tasks (Wu et al., 2023; Phan et al., 2024). We partition the dataset into training, validation, and test sets with a ratio of $0.6/0.2/0.2$, respectively.

SIIM-ACR Pneumothorax (sii, 2019). SIIM-ACR Pneumothorax contains 12,954 chest X-ray images, along with image-level pneumothorax annotations and pixel-level segmentation masks where pneumothorax is present. Like the RSNA Pneumonia dataset, it can be used for both classification and segmentation tasks. We divide the dataset into training, validation, and test sets with a ratio of $0.6/0.2/0.2$.

CheXpert (Irvin et al., 2019). CheXpert contains 224,316 chest X-ray images from 65,240 patients, collected by Stanford Hospital. The official test set includes images from 500 patients, annotated through consensus by five board-certified radiologists. We evaluated all disease categories in this test dataset.

COVIDx CXR-2 (Pavlova et al., 2022) and COVID Rural (Desai et al., 2020). The COVIDx CXR-2 and COVID Rural are designed for evaluating COVID-19 diagnosis. COVIDx CXR-2 (Pavlova et al., 2022) consists of 29,986 images from 16,648 COVID-19 patients, each labeled with a classification tag. The dataset is split into training, validation, and test sets with a ratio of $0.7/0.2/0.1$, used for evaluating classification performance. The COVID Rural dataset contains over 200 chest X-ray images with annotated segmentation masks, used for the COVID-19 segmentation task. This dataset is partitioned into training, validation, and test sets with a ratio of $0.6/0.2/0.2$.

C Medical Category Semantic Vector Library

We draw inspiration from the work of MedKLIP (Wu et al., 2023) and incorporate 75 frequently occurring medical entities from clinical reports as input to our model. By designing prompts, we stimulate the model's intrinsic medical knowledge, enabling it to infer the semantic representations of various entity categories. The resulting semantic descriptions of these 75 medical entities are presented in table 7.

Furthermore, to achieve a more precise representation of major disease categories, we construct a dedicated disease semantic vector library, which facilitates a more nuanced understanding of disease-related semantics. The generated disease descriptions are detailed in table 6.

D Implementation Details

Unless otherwise specified, we use LLaVA-Med (Li et al., 2024a) as the foundational MLLM $\mathcal{F}$. We employ the LoRA strategy for parameter-efficient fine-tuning, with training managed via the DeepSpeed engine.

For optimization, we utilize the AdamW optimizer with a learning rate of 2e-5 and no weight decay. A cosine learning rate decay schedule is applied, with 3% of the total training steps allocated for warm-up. The number of special tokens for images $< \mathit{ImgCls} >$ is set to 4, while the number of special tokens for text $< \mathit{TxtCls} >$ is set to 8. The temperature hyperparameter τ is configured as 0.05, and the loss weight coefficient λ is set to 0.5. Furthermore, the batch size per GPU is set to 64.

E Use of LLMs

This paper employed large language models (i.e., ChatGPT, Claude) solely for language editing and polishing purposes, including but not limited to grammar checking, expression optimization, and text refinement. All core research content, including experimental design, data analysis, and conclusion derivation, was carried out independently by the authors. The authors take full responsibility for the entire content of this paper and have thoroughly verified and validated all AI-assisted modifications.

Table 6: Semantic Descriptions of 14 Medical Disease Categories

Disease	Description
Fibrosis	Fibrosis refers to excessive deposition of collagen and extracellular matrix during abnormal tissue repair after inflammation or injury, leading to the replacement of normal lung tissue with reticular or band-like high-density shadows, commonly seen in the lower and peripheral lungs. Imaging may show honeycombing and traction bronchiectasis. Clinically, patients often present with progressive dyspnea, dry cough, and reduced exercise tolerance.
Edema	Pulmonary edema refers to the abnormal accumulation of fluid in the pulmonary interstitium and alveoli, usually caused by cardiogenic or non-cardiogenic factors. Imaging shows patchy or 'bat-wing' distributed heterogeneous high-density shadows in the middle or entire lung, often accompanied by Kerley lines and cardiac enlargement. Clinically, patients typically experience acute dyspnea, cough, cyanosis, and bilateral lung crackles.
Pneumothorax	Pneumothorax refers to the presence of air in the pleural cavity, leading to partial or complete lung collapse. Imaging typically shows a low-density black air space along the pleura, with a clear demarcation from the normal lung tissue, along with lung collapse. In tension pneumothorax, mediastinal shift may occur. Clinically, patients often present with sudden unilateral chest pain, dyspnea, and decreased breath sounds, sometimes accompanied by subcutaneous emphysema.
Cardiomegaly	Cardiomegaly refers to the enlargement of the heart due to hypertension, cardiomyopathy, or valvular disease, causing chamber dilation or wall thickening. Imaging shows significant cardiac enlargement with an expanded and smooth contour, often marked by an increased cardiothoracic ratio, potentially accompanied by pulmonary congestion and bronchial congestion. Clinically, patients may experience reduced exercise tolerance, dyspnea, lower limb edema, and arrhythmias.
Atelectasis	Atelectasis refers to the collapse of part or all of the lung tissue due to airway obstruction, external thoracic pressure, or intrapulmonary pathology. Imaging shows increased local lung density, volume reduction, bronchial displacement, and visceral pleural traction, commonly affecting the lower lobes. Clinically, patients may exhibit rapid shallow breathing, localized decreased or absent breath sounds, and a history of recent surgery or inadequate airway clearance.
Nodule	A lung nodule is a localized lesion with a diameter of less than 3 cm. Imaging typically shows a round or oval localized density, with either well-defined or spiculated edges. Some nodules may contain calcifications or low-density necrotic areas. Clinically, most patients are asymptomatic, but growing or malignant nodules may present with cough and hemoptysis.
Emphysema	Emphysema is a chronic obstructive pulmonary disease caused by the permanent destruction of alveolar walls and airspace enlargement. Imaging shows scattered or diffuse low-density areas in both lungs, reduced lung markings, often with bullae or cystic lesions, a flattened diaphragm, and hyperinflated lungs. Clinically, patients typically have a history of chronic cough, sputum production, and progressive dyspnea, often associated with smoking or long-term occupational exposure.
No Finding	No finding refers to the absence of radiographic abnormalities detected in the chest X-ray.

Table 6: Semantic Descriptions of 14 Medical Disease Categories

Disease	**Description**
Mass	A mass refers to an abnormal localized tissue overgrowth. Imaging shows a focal high-density lesion, which may have regular or irregular shapes with spiculated margins, often accompanied by internal necrosis, calcification, or hemorrhage. Surrounding features may include bronchial distortion or lymphadenopathy. Clinically, patients may present with cough, weight loss, or hemoptysis, requiring further pathological examination.
Pleural Thickening	Pleural thickening refers to fibrotic or calcified pleural changes due to chronic inflammation, infection, or asbestos exposure. Imaging shows localized or diffuse thickening along the pleural surface, appearing as streaky or patchy high-density shadows, sometimes with nodular changes. Clinically, patients may be asymptomatic, but a history of pleuritis or exposure to harmful substances is often present.
Effusion	Pleural effusion refers to the abnormal accumulation of fluid in the pleural cavity, which may be caused by infection, heart failure, malignancy, or other inflammatory diseases. Typically seen in the lower lung fields and posterior chest cavity, imaging shows a homogeneous or layered fluid density with a clear meniscus sign, with CT revealing low-density regions. Severe effusion may cause lung compression or bronchial displacement. Clinically, patients may present with dyspnea, chest pain, and cough, with physical signs of reduced breath sounds, dull percussion, and abnormal auscultation.
Infiltration	Infiltration refers to localized or diffuse high-density changes in lung tissue due to inflammation, infection, or malignancy. Imaging typically shows patchy or ill-defined high-density areas, sometimes with a ground-glass appearance or consolidation, occasionally accompanied by air bronchograms or bronchial wall thickening. Clinically, patients may present with cough, fever, dyspnea, and fatigue, often with elevated inflammatory markers.
Pneumonia	Pneumonia refers to lung parenchyma inflammation caused by bacteria, viruses, fungi, or other microorganisms, leading to alveolar filling with inflammatory exudates. Imaging shows localized or patchy consolidation with irregular margins, often accompanied by air bronchograms, pleural reaction, and mild pleural effusion. Clinically, patients present with fever, cough, sputum production, chest pain, and fatigue, with elevated white blood cell counts and inflammatory markers.
Consolidation	Consolidation refers to the complete filling of alveolar spaces with liquid, pus, blood, or cellular material, replacing the normal air content. Imaging shows homogeneous, dense, well-defined opacities, often with air bronchograms and pleural reactions, sometimes with minimal pleural effusion. Clinically, patients often have fever, cough, sputum production, chest pain, and dyspnea, with significantly elevated inflammatory markers.

Table 7: Semantic Descriptions of 75 Medical Categories

Disease	**Description**
normal	Indicates that the structure appears within standard parameters without signs of pathology.
clear	The imaging reveals no obscuring abnormalities, ensuring clear visualization of the structure.
sharp	Boundaries are precisely defined, accentuating the distinct separation between tissues.
sharply	The structure is rendered with exceptional clarity, facilitating detailed evaluation.
unremarkable	No significant deviations or abnormalities are observed in the examined area.
intact	The structure remains whole and undamaged, with no disruption detected.
stable	The tissue exhibits consistent appearance over time without progressive changes.
free	Presence of extraluminal air in unexpected locations, possibly indicating a perforation.
effusion	Accumulation of fluid between the pleural layers, often reflecting an underlying pathology.
opacity	An area of increased radiodensity that obscures normal lung markings, suggesting fluid or tissue replacement.
pneumothorax	Air present in the pleural space that may lead to partial or complete lung collapse.
edema	Diffuse fluid accumulation within lung tissue, frequently associated with cardiac or inflammatory issues.
atelectasis	Collapse of lung segments resulting in volume loss and increased local density.
tube	A medical tube visible on imaging, such as for drainage or airway management.
consolidation	Region where alveolar air is replaced by fluid or cells, producing homogeneous density.
process	Denotes an active pathological condition altering the tissue's normal appearance.
abnormality	A generic term for any deviation from normal structure suggestive of disease.
enlarge	Indicates that a structure appears larger than typical normal values.
tip	The distal or pointed end of a structure or medical device.
low	Underinflation of the lungs, often implying a restrictive process.
pneumonia	Inflammatory infection of lung parenchyma, typically showing consolidation and air bronchograms.
line	A linear structure that may represent a fissure, pleural interface, or artifact.
congestion	Increased blood or fluid accumulation in tissues, often indicating impaired circulation.
catheter	A slender, flexible tube inserted for drainage or medication delivery, visible in imaging.
cardiomegaly	An enlarged cardiac silhouette, frequently associated with chronic heart conditions.
fracture	A break or discontinuity in bone structure evident on radiographs.
air	Regions of radiolucency indicating the presence of gaseous content.
tortuous	Describes a vessel or structure exhibiting excessive curvature or winding.
lead	The foremost or guiding portion of a device or anatomical feature.
disease	A general term for any pathological process affecting normal tissue function.

Table 7: Semantic Descriptions of 75 Medical Categories

Disease	**Description**
calcification	Deposition of calcium salts within tissue, appearing as bright foci on imaging.
prominence	An area that appears more pronounced than surrounding tissues, suggesting an increase in size or density.
device	Any implanted or externally applied apparatus used for diagnostic or therapeutic purposes.
engorgement	Excessive filling of vessels or tissues with blood, leading to a swollen appearance.
picc	A long, thin catheter introduced via a peripheral vein and advanced into the central circulation for long-term therapy.
clip	A small metallic or plastic fastener used during surgery to secure tissues or vessels.
elevation	An upward displacement or raised position of an anatomical structure relative to its usual location.
expand	Describes a structure that appears dilated or increased in volume.
nodule	A small, rounded lesion typically less than 3 cm in diameter that can be benign or malignant.
wire	A thin, flexible metallic strand often used in surgical fixation or as part of medical devices.
fluid	The presence of liquid within tissues or cavities, altering the normal radiographic appearance.
degenerative	Changes in tissue structure resulting from chronic wear, aging, or repeated stress.
pacemaker	An implanted device that regulates heart rhythm, visible through its leads and generator.
thicken	Describes a structure that appears denser or more layered, possibly due to fibrotic changes.
marking	Visible patterns or lines that may represent vascular or connective tissue features.
scar	Fibrotic tissue that replaces normal parenchyma following injury, typically seen as an irregular opacity.
hyperinflate	Denotes lungs that are over-expanded, often with increased radiolucency and flattened diaphragms.
blunt	Loss of sharp definition in anatomical borders, leading to a less distinct appearance.
loss	Indicates a reduction or absence of normal tissue volume or density.
widen	Suggests that a structure or space is broader than the standard measurement.
collapse	A significant reduction or complete loss of volume in lung tissue due to obstruction or injury.
density	Reflects the compactness of a tissue, with higher density appearing whiter on radiographs.
emphysema	A chronic condition marked by alveolar wall destruction and abnormal enlargement of air spaces.
aerate	Indicates that the lung tissue is adequately filled with air, supporting effective gas exchange.
mass	A malignant tumor arising from lung tissue, typically presenting as an irregular mass with possible cavitation.
crowd	Compaction of airways and vessels, often due to volume loss or infiltrative processes.
infiltrate	Diffuse or patchy opacities in the lung that suggest inflammation, infection, or neoplastic involvement.
obscure	Describes anatomical structures that are not clearly visualized, often due to overlapping tissues or technical factors.

Table 7: Semantic Descriptions of 75 Medical Categories

Disease	Description
deformity	An abnormal shape or structure resulting from congenital anomalies, trauma, or disease progression.
hernia	The protrusion of an organ or tissue through an abnormal opening in the surrounding structure.
drainage	The process or presence of fluid removal from a body cavity, often via an inserted tube.
distention	Abnormal expansion or swelling of a structure due to accumulation of fluid or gas.
shift	Displacement of anatomical structures from their usual positions, indicating mass effect or volume change.
stent	A small mesh tube used to maintain the patency of a vessel or duct.
pressure	The force exerted per unit area by fluids or tissues, which can influence organ function.
lesion	Any abnormal area of tissue that deviates from the standard architecture, potentially indicative of pathology.
finding	A generic term for an observed abnormality or noteworthy feature on imaging.
borderline	The heart appears at the upper limit of normal size, without clear evidence of enlargement.
hardware	Any implanted or externally attached device used for diagnostic, therapeutic, or supportive purposes.
dilation	The widening or expansion of a hollow structure, often reflecting increased internal pressure.
chf	A clinical syndrome characterized by the heart's reduced pumping ability, leading to systemic fluid accumulation.
redistribution	A shift in the normal pattern of blood or air distribution within the lungs, often due to altered hemodynamics.
aspiration	Inhalation of foreign material into the airways, potentially leading to inflammatory or infectious complications.
rare diseases	Conditions that occur infrequently in the population and often require specialized diagnostic and management approaches.
Covid-19	An infectious disease caused by the SARS-CoV-2 virus, with a broad spectrum of respiratory and systemic manifestations.