Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models

Songtao Jiang[*1], Tuo Zheng[*1,2], Yan Zhang[3], Yeying Jin[3], Li Yuan[4] and Zuozhu Liu[†1]

[1]ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University
[2]ChohoTech Inc., Hangzhou
[3]National University of Singapore
[4]Peking University

Abstract

Recent advancements in general-purpose or domain-specific multimodal large language models (LLMs) have witnessed remarkable progress for medical decision-making. However, they are designated for specific classification or generative tasks, and require model training or finetuning on large-scale datasets with sizeable parameters and tremendous computing, hindering their clinical utility across diverse resource-constrained scenarios in practice. In this paper, we propose a novel and lightweight framework Med-MoE (Mixture-of-Experts) that tackles both discriminative and generative multimodal medical tasks. The learning of Med-MoE consists of three steps: multimodal medical alignment, instruction tuning and routing, and domain-specific MoE tuning. After aligning multimodal medical images with LLM tokens, we then enable the model for different multimodal medical tasks with instruction tuning, together with a trainable router tailored for expert selection across input modalities. Finally, the model is tuned by integrating the router with multiple domain-specific experts, which are selectively activated and further empowered by meta expert. Comprehensive experiments on both open- and close-end medical question answering (Med-VQA) and image classification tasks across datasets such as VQA-RAD, SLAKE and Path-VQA demonstrate that our model can achieve performance superior to or on par with state-of-the-art baselines, while only requiring approximately 30%-50% of activated model parameters. Extensive analysis and ablations corroborate the effectiveness and practical utility of our method. Our code is released at https://github.com/jiangsongtao/Med-MoE.

* Equal contribution, part of this work was done when Tuo Zheng was an intern at ChohoTech.
† Corresponding author: zuozhuliu@intl.zju.edu.cn

1 Introduction

Creating systems with human-level multimodal understanding is essential for medical decision-making (Miao et al., 2022; Goyal et al., 2016; de Faria et al., 2023; Antol et al., 2015). Recent progress on Multimodal Large Language Models (MLLMs) such as LLaVA (Liu et al., 2024a), MiniGPT4-V2 (Chen et al., 2023), CogVLM (Wang et al., 2023) have demonstrated great performance across multimodal tasks, however, they are less effective in the medical domain as they are usually trained with web contents which differ significantly from the medical data. Domain-specific models such as Med-Flamingo (Moor et al., 2023), Med-PaLM M (Singhal et al., 2023), and LLaVA-Med (Li et al., 2024a) exhibit promising results across various medical tasks, such as medical visual question-answering (Med-VQA), by training with medical domain data. However, these models are usually tailed for certain kinds of tasks, such as close- or open-end VQA, while in practice, medical MLLMs need to handle both discriminative and generative tasks to provide more reliable and interpretable decisions. Moreover, existing models are usually obtained with heavy LLMs with sizeable parameters, such as LLama(7B) in LLaVA-Med, leading to high training and inference costs and hindering its practical utility to broad clinical practitioners.

It is appealing but challenging to build lightweight yet effective medical MLLMs for multimodal decision-making (Petersson et al., 2022; Kelly et al., 2019; Liao et al., 2024). Recent research shows that scaling up the quantity or quality of training data, as well as increasing the size of the model, can result in enhanced performance (Gao et al., 2024; Shi et al., 2024; Xue et al., 2024; Shen et al., 2023; Lu et al., 2023). However, training and deploying these models also demand substantial computational resources, rendering them
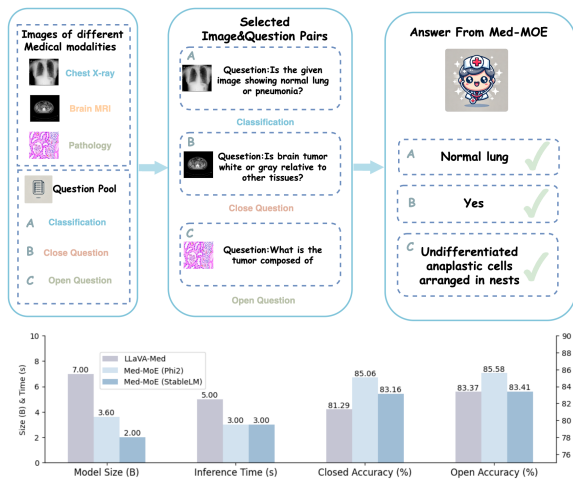
Figure 1: **Upper:** This figure showcases our model's capability in addressing three primary types of Medical VQA challenges and image classification tasks. **Lower:** Comparison between Med-MoE and LLaVA-Med, emphasizing Med-MoE's advantages in inference speed, model size, and its superior performance.

less appealing for numerous clinical practitioners who may lack sufficient computing power (Lu et al., 2023; Crawford, 2021; Thompson et al., 2020). For instance, many institutions may not possess powerful GPUs such as NVIDIA A100 cards to tune LLama-7B model family, e.g., LLaVA-Med. Moreover, the medical data differs drastically from web contents, and its inherent multi-modality, such as imaging from CT, MRI, X-ray and pathology, presents additional challenges to develop effective yet lightweight medical MLLMs (Acosta et al., 2022; Xu et al., 2024b). This task becomes even more difficult when considering the requirements for reliability and interpretability in decision-making (Salahuddin et al., 2022; Vellido, 2020).

Recent work explores cost-effective training of light-weight LLMs by architecture design, training procedure or hardware optimization etc (Dubiel et al., 2024; Zhao et al., 2024; Hu et al., 2024; Zhou et al., 2024). Among these techniques, the Mixture-of-Expert (MoE) strategy has shown great potential for general-purpose training (Chen et al., 2022; He et al., 2021; Jacobs et al., 1991; Eigen et al., 2013), e.g., the Mixtral family employs sparse MoEs to achieve competing performance with LLama-70B with only 12.9B active parameters (Jiang et al., 2024); the MoE-LLaVA propose a MoE-based sparse large VLM framework with novel training strategies (Lin et al., 2024). By combining multiple small-scale sub-modules, i.e., experts, and activating only the top-k relevant experts for each task,

the MoE model can achieve good performance with much less computing cost. Despite these successes in general domains, current MoEs often overlook the specialization and synergy of experts required in medical contexts, and their applicability in the medical domain remains unexplored.

In this paper, we propose a lightweight and effective framework Med-MoE for multimodal generative or discriminative Med-VQA and classification tasks. Our Med-MoE incorporates multiple domain-specific experts along with global meta expert, emulating the workflow in hospitals where various departments collaborate together for disease diagnosis. In particular, the Med-MoE takes lightweight LLMs with smaller sizes of parameters as the base model of experts, which are first trained with medical image and caption pairs to align visual and textual modalities. Afterward, the model is trained with medical instruction following datasets to better perform multimodal medical tasks. Meanwhile, a router is trained to identify different medical image modalities, enabling better selection across multiple domain-specific experts during decision-making. Inspired by the well-known ResNet (He et al., 2016) architecture and the Multi-Disciplinary Team (MDT) diagnosis mechanism in clinics, we propose to add an additional meta expert in the shortcut, as shown in Figure 2, which captures global medical information to assist the specified expert for better performance. During inference, only the meta expert and the selected experts are activated, leading to a lightweight model with only a small portion of activated parameters.

The Med-MoE consistently demonstrates significant performance improvements across diverse medical datasets, encompassing both open- and close-end Med-VQA and medical image classification tasks in VQA-RAD, SLAKE, PathVQA, PneumoniaMNIST, and OrganCMNIST. Comprehensive experiments show that our Med-MoEs, which are constructed with two small-scale LLMs, i.e., Phi2 (2.7B) (Abdin et al., 2024) and StableLM (1.7B) (Bellagente et al., 2024), can attain performance superior to or on par with the state-of-the-art LLaVA-Med (7B) model, with only 2.0-3.6B activated parameters. Extensive ablations and analysis demonstrate the efficacy of our Med-MoE in advancing multimodal medical tasks and highlight its potential to enhance outcomes in resource-limited healthcare settings.
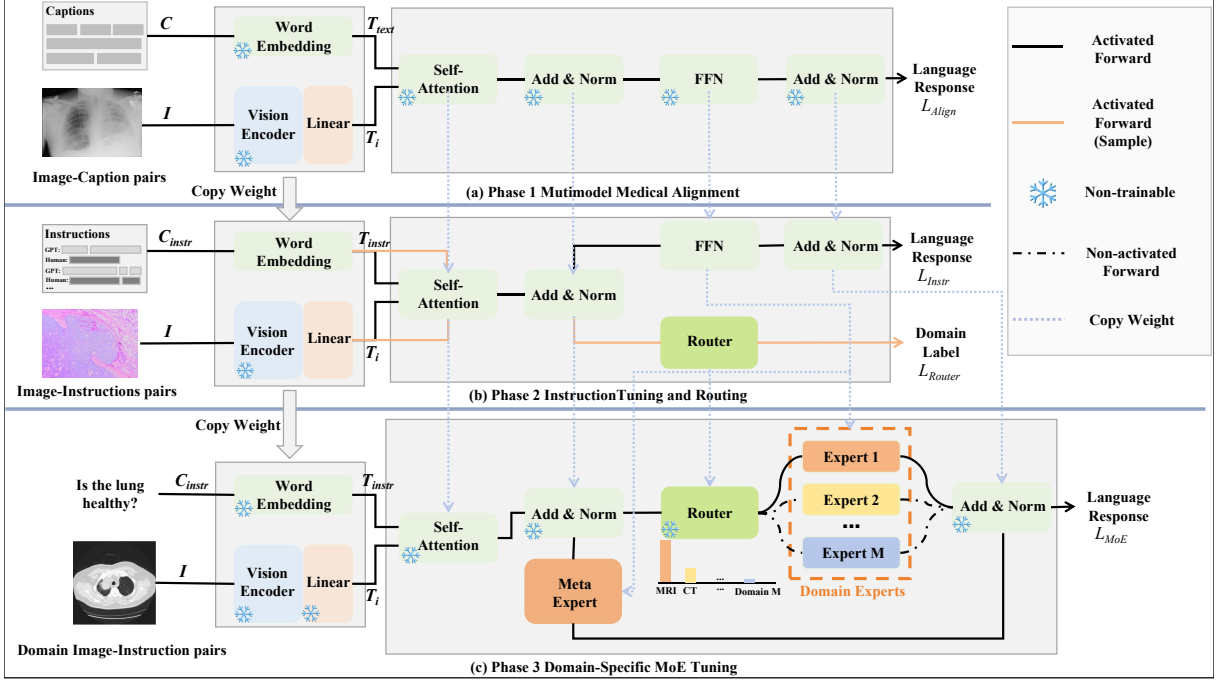
Figure 2: The framework of Med-MoE with three phases.

2 Methods

The training of Med-MoE includes three phases, as illustrated in Figure 2. First, we perform multi-modal medical alignment to help the LLM to comprehend medical images by leveraging the vision encoder's image tokens. Next, we conduct instruction tuning to enable the model to execute various medical tasks and enhance its instruction-following ability. Meanwhile, a router is trained with a small amount of labeled data to characterize the input modality. Finally, we perform domain-specific MoE tuning by replacing the model's FFN with sparsely activated experts, where a meta-expert is always activated to capture global information.

2.1 Phase 1: Multimodal Medical Alignment

In this phase, we train only the MLP following the vision encoder to achieve modality alignment. We use the same vision encoder as LLaVA-Med, specifically the pretrained CLIP-ViT-Large-Patch14 released by OpenAI (Radford et al., 2021). We align visual and textual modalities by curating a dataset of medical images I paired with corresponding captions C. The images I are fed into a vision encoder E_v to produce image tokens T_i ($T_i = E_v(I_i)$), and the captions C are tokenized into text tokens T_{text} ($T_{\text{text}} = \text{Tokenizer}(C_i)$). The concatenated tokens T_{comb} are fed into the LLM, which is trained to generate the continuation of text tokens, minimizing

the self-supervised loss:

$$\mathcal{L}_{\text{Align}} = -\sum_{i=1}^{N} \log p\left(T_{\text{text}}^{[P+i]} \mid T_{\text{comb}}, T_{\text{text}}^{[:i-1]}\right). \tag{1}$$

2.2 Phase 2: Instruction Tuning and Routing

This phase aims to enhance the model's ability to follow complex medical instructions to perform various multimodal tasks and train a router for expert selection in the next phase. The instruction tokens T_{instr} and image tokens T_i are concatenated into T_{comb} and fed into the LLM. The model is trained using a dataset of medical queries and responses to generate accurate responses, minimizing the loss:

$$\mathcal{L}_{\text{Instr}} = -\sum_{i=1}^{N} \log p\left(T_{\text{resp}}^{[P+i]} \mid T_{\text{comb}}, T_{\text{resp}}^{[:i-1]}\right). \tag{2}$$

We also train a router to predict the input modality using a small subset of data with the loss:

$$\mathcal{L}_{\text{Router}} = -\sum_{i=1}^{M} y_i \log p(y_i \mid T_{\text{comb}}), \tag{3}$$

where y_i is the true label of the input image modality, i.e., CT, MRI, Pathology and X-ray, etc. In this phase, the vision encoder is frozen, while all other components are trained. Our router uses a single-layer MLP structure, which has been widely used in previous research (Jacobs et al., 1991; Fedus et al.,

2022). To enable the router to activate different experts based on input modality, we used a small subset of data, selecting 200 labeled examples from each modality for router training. After completing the second stage of training, the router predicts the modality of the input data and calculates the cross-entropy loss function with the predicted modality and the true modality labels.

2.3 Phase 3: Domain-Specific MoE Tuning

Finally, we replace the LLM's FFN with MoE(mixture-of-experts) architecture. The router, trained in phase 2, assigns inputs to specific experts, while a meta-expert is always activated to capture global information. The MoE layer's output is a weighted combination of the experts' outputs. The domain-specific experts and the meta-expert are initialized with the FFN weights from the model trained in phase 2. In this phase, the router from phase 2 is used and frozen, so only the domain-specific experts and the meta-expert are trained.

$$\mathbf{O}_{\text{MoE}} = \sum_{i=1}^{K} G_i E_i + E_{\text{meta}}, \qquad (4)$$

where G_i is the gating function provided by the router, E_i are the domain-specific experts, and E_{meta} is the meta-expert. The training loss is:

$$\mathcal{L}_{\text{MoE}} = -\sum_{i=1}^{N} \log p\left(T_{\text{resp}}^{[P+i]} \mid \mathbf{O}_{\text{MoE}}, T_{\text{resp}}^{[:i-1]}\right). \quad (5)$$

In the end, the model is fine-tuned for specific medical domains, leveraging expert knowledge to provide highly accurate and relevant responses across the open- and close-end and classification tasks.

3 Experiment

3.1 Experiment Settings

Dataset: We utilize well-organized datasets provided by LLaVA-Med (Li et al., 2024a) for alignment and instruction tuning in phase 1&2, see details in Supplementary Figure 13. In MoE-tuning phase, we employ VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020) with open- and close-end QA pairs for Med-VQA tuning and evaluation. For classification task, we use the PneumoniaMNIST and OrganCMNIST from (Yang et al., 2023). Detailed information and examples of Med-MoE's responses are shown in Supplementary.

Evaluation Metrics: We employ the accuracy for closed-set questions and recall for open-set

questions, being consistent with existing work like LLaVA-Med for a fair comparison. In Table 2, we also evaluate the exact match and BLEU scores for comprehensive evaluation.

Experiment Setup: We select two small LLMs, i.e., StableLM (1.7B) and Phi2 (2.7B), as the base model, see Figure 14. The size of activated parameters in resulting Med-MoEs are 2.0B (Med-MoE StableLM) and 3.6B (Med-MoE Phi2), with additional parameters from domain-specific experts and meta-experts. To ensure a fair comparison with LLaVA-Med, we also train a LLaVA-Med model using the Phi2 (2.7B) backbone. This allows us to compare the performance under the same LLM backbone. We also investigate the versatility of our method by combining it with other cost-efficient approaches, such as LoRA-based methods. Our experimental hyperparameters are shown in Supplementary Table 12.

Baselines: We compare our method with a diverse set of baselines: (1) **CLIP-based methods**, such as BiomedCLIP and CLIP-ViT (Zhang et al., 2023b; Eslami et al., 2023), which are state-of-the-art in this category but are limited by their reliance on candidate words for answering questions in open settings; (2) **OFA (One for All)-based models**, like the recent BiomedGPT (Zhang et al., 2023a), which leverage generative multimodal pretraining and have shown promising performance in the medical field, but their lack of multi-turn dialogue capability, due to not being LLM-based, restricts their usage in clinical practice; (3) **MLLM-based models**, including Med-Flamingo and the state-of-the-art LLaVA-Med, which, despite their impressive VQA performance, have large parameter sizes (7B and above) that hinder their applicability in real-world clinical settings. In classification tasks, we compare with ViT-based methods and the latest Med-Mamba (Yue and Li, 2024). Notably, our Med-MoE, an MLLM-based method, offers multi-turn dialogue capabilities for open VQA settings which are not present in traditional methods, while exhibiting effective training/inference and competing performance.

3.2 Main Results

Zero-shot Performance on Med-VQA tasks: Our models exhibit notable improvements in zero-shot performance across various medical VQA tasks. The Med-MoE (Phi2) model boosts scores by approximately 1.4% in VQA-RAD Open, 2.6% in VQA-RAD Closed, 5.3% in SLAKE Open, and

Method	VQA-RAD		SLAKE		PathVQA		Act.
	Open	Closed	Open	Closed	Open	Closed	
Representative & SoTA methods with numbers reported in the literature (Non-MLLM Based Methods)							
VL Encoder–Decoder (Bazi et al., 2023)	-	82.47	-	-	-	85.61	-
Q2ATransformer (Liu et al., 2023)	-	81.20	-	-	54.85	88.85	-
Prefix T. Medical LM (van Sonsbeek et al., 2023)	-	-	-	82.01	-	87.00	-
PubMedCLIP (Eslami et al., 2023)	-	80.00	-	82.50	-	-	-
BiomedCLIP (Zhang et al., 2023b)	-	79.80	-	89.70	-	-	-
M2I2 (Li et al., 2022)	-	83.50	-	91.10	-	88.00	-
BiomedGPT-S (Zhang et al., 2023a)	13.40	57.80	66.50	73.30	10.70	84.20	-
BiomedGPT-M (Zhang et al., 2023a)	53.60	65.07	78.30	86.80	12.5	85.70	-
CLIP-ViT w/ GPT2-XL	-	-	84.30	82.10	40.0	87.00	-
Supervised finetuning results (MLLM Based Methods)							
LLaVA	50.00	65.07	78.18	63.22	7.74	63.20	7B
LLaVA-Med (LLama7B)	<u>61.52</u>	**84.19**	83.08	<u>85.34</u>	<u>37.95</u>	91.21	7B
LLaVA-Med (Vicuna7B)	**64.39**	81.98	<u>84.71</u>	83.17	**38.87**	<u>91.65</u>	7B
LLaVA-Med (Phi2.7B)	54.83	81.35	81.29	83.29	31.73	90.17	2.7B
Med-MoE (Phi2)	58.55	<u>82.72</u>	**85.06**	**85.58**	34.74	**91.98**	3.6B
Med-MoE (StableLM)	50.08	80.07	83.16	83.41	33.79	91.30	2.0B
Zero-shot results							
LLaVA-Med (LLama7B)	<u>36.23</u>	60.16	<u>41.72</u>	47.60	**10.86**	59.75	-
Med-MoE (Phi2)	**36.73**	<u>61.75</u>	**43.93**	**56.97**	6.94	<u>66.46</u>	-
Med-MoE (StableLM)	28.02	**66.91**	40.63	<u>52.64</u>	<u>9.40</u>	**69.09**	-

Table 1: Performance on Med-VQA tasks. **Bold** denotes the best performance; <u>underlined</u> denotes the second-best.

Method		VQA-RAD			SLAKE			PathVQA		
		EMS	R	BS	EMS	R	BS	EMS	R	BS
LLaVA-Med	7B	<u>58.33</u>	**61.52**	<u>54.13</u>	<u>82.83</u>	83.08	81.69	**37.95**	**36.86**	<u>32.89</u>
Med-Flamingo(Few-Shot) (Moor et al., 2023)	9B	20.00	-	-	-	-	-	31.00	-	-
PaLM-E (Tu et al., 2024)	84B	-	-	**59.19**	-	-	52.65	-	-	**54.92**
Med-MoE (Phi2)	3.6B	**59.69**	<u>58.55</u>	52.95	**84.46**	**85.06**	**83.16**	<u>34.37</u>	<u>34.74</u>	32.85
Med-MoE (StableLM)	2.0B	52.53	50.08	45.67	82.44	<u>83.16</u>	81.53	33.60	33.79	32.67

Table 2: Detailed comparison regarding more metrics (Supplementary A) in Open settings.

Methods	PneumoniaMNIST	OrganCMNIST
Med-Mamba (Yue and Li, 2024)	<u>91.2</u>	**92.4**
AutoKeras (Jin et al., 2019)	87.8	87.9
BiomedGPT	90.8	88.9
Med-MoE (StableLM)	89.3	88.6
Med-MoE (Phi2)	**91.4**	<u>89.9</u>

Table 3: Image classification accuracy comparison.

9.4% in SLAKE Closed compared to LLaVA-Med (LLama7B). The Med-MoE (StableLM) variant achieves around 6.8% higher in VQA-RAD Closed, 5.0% in SLAKE Closed, and 9.3% in PathVQA Closed, demonstrating robust performance. These results highlight the superior effectiveness of Med-MoE models in zero-shot settings.

Comparison with SOTA Methods on Med-VQA: Overall, Med-MoE can achieve superior or competing performance with the best-performing LLaVA-Med (7B) with only 2.0B or 3.6B activated parameters. In particular, Med-MoE (Phi2) surpasses the best LLaVA-Med variants in SLAKE Open (85.06), SLAKE Closed (85.58), and PathVQA Closed (91.98), and shows competing performance on the rest tasks. Med-MoE (StableLM) also exhibits better performance than the LLaVA-Med (Phi-2.7B) in most scenarios, with only 2.0B activated parameters. Its performance is also on par with LLaVA-Med (7B) in many scenarios, with even better performance in SLAKE Closed. These results highlight the effectiveness and strong potential of Med-MoE to establish new benchmarks across various datasets and tasks.

Results on Medical Image Classification: In contrast to most existing LLM-based work, e.g., LLaVA-Med, which only evaluates the performance on Med-VQA, we further evaluate Med-MoE on classification tasks for comprehensive analysis. As shown in Table 3, Med-MoE (Phi2) achieves 91.4% accuracy on PneumoniaMNIST, outperforming BiomedGPT and Med-Mamba. It

also showcases the second-best performance on Or-ganCMNIST, closely following Med-Mamba. The performance of Med-MoE (StableLM) is a little bit worse. Overall, the classification performance of Med-MoE is quite promising, while its performance might be boosted if more relevant data rather than image-caption pairs could be used for model alignment and tuning in the initial phases.

4 Ablation and Analysis

Method	SFT	MoE-Tuning
VQA-RAD (Open)	54.83	**58.55 (+3.72)**
VQA-RAD (Closed)	81.35	**82.72 (+1.37)**
SLAKE (Open)	81.29	**85.06 (+3.77)**
SLAKE (Closed)	83.29	**85.58 (+2.29)**
PathVQA (Open)	31.73	**34.74 (+3.01)**
PathVQA (Closed)	90.17	**91.98 (+1.81)**

Table 4: Comparison of SFT and MoE Tuning.

Method	No Meta Expert	With Meta Expert
VQA-RAD (Open)	54.37	**58.55 (+4.18)**
VQA-RAD (Closed)	81.42	**82.72 (+1.30)**
SLAKE (Open)	81.54	**85.06 (+3.52)**
SLAKE (Closed)	82.45	**85.58 (+3.13)**
PathVQA (Open)	32.12	**34.74 (+2.62)**
PathVQA (Closed)	90.19	**91.98 (+1.79)**

Table 5: Ablation on the meta expert.

Method	Learned Router	Router (Ours)
VQA-RAD (Open)	56.33	**58.55 (+2.22)**
VQA-RAD (Closed)	82.19	**82.72 (+0.53)**
SLAKE (Open)	82.75	**85.06 (+2.31)**
SLAKE (Closed)	84.59	**85.58 (+0.99)**
PathVQA (Open)	33.40	**34.74 (+1.34)**
PathVQA (Closed)	91.19	**91.98 (+0.79)**

Table 6: Ablation on the routing mechanism.

Ablation of Router: We evaluate the effectiveness of our routing mechanism compared to the general MoE routing mechanism across the Med-VQA datasets. The Learned Router refers to the router in MoE that is trained during the MoE tuning phase, while our routers are pretrained with modality information, leading to more efficient and accurate routing. Results in Table 6 for the Phi2.7B model show that our router achieves consistent improvements. The improvements in open settings are even more evident than those in closed settings, demonstrating the effectiveness of our routing mechanism in more challenging open scenarios.

Ablation of Meta Expert: We evaluate the impact of the meta expert with ablation results with the Phi2.7B model in Table 5. We can notice that the meta expert can bring consistent and significant improvements over all Med-VQA settings, with improvements of 1.30-4.18%. These consistent improvements underscore the critical role of the meta expert in enhancing the model's ability to cope with various multimodal medical tasks.

Ablation of Domain-Specific MoE-Tuning: To assess the benefits of the MoE-Tuning over traditional Supervised Fine-Tuning (SFT), we conduct an ablation study with the Phi2.7B model. Results across three Med-VQA datasets (Table 4) demonstrate that the MoE-Tuning can lead to better performance. These results demonstrate the effectiveness of our MoE architecture, giving rise to better performance than tuning a dense FFN.

Comparison with LoRA-based Methods: We further investigate the compatibility of our methods with other lightweight techniques, i.e., LoRA. As shown in Table 7, by integrating with LoRA, Med-MoE exhibits much less performance degradation compared to LLaVA-Med. For instance, in the SLAKE Closed setting, Med-MoE (Phi2) with LoRA exhibits only a 0.49% performance drop, whereas LLaVA-Med with LoRA experiences 1.97% degradation. Furthermore, we observe that LoRA reduces GPU memory usage during training, and our Med-MoE requires fewer activated parameters during inference. Consequently, the integration of LoRA with Med-MoE achieves lightweight learning in terms of both training and inference. These findings indicate that Med-MoE presents an appealing practical choice for medical tasks, delivering promising performance at significantly lower computational costs.

Effect of Architectures and Training Data of Router: Figure 3 investigates the effectiveness of different MLP structures in the router. We can notice that complicated MLPs, e.g., using 3 MLP layers, might not give rise to consistent improvements and may even lead to overfitting. As shown in Figure 3, a simple MLP with 1 or 2 layers can learn good embeddings of the input modality, resulting in clusters with clear boundaries, as well as high accuracy and Silhouette score. Figure 4 confirms effective modality differentiation with well-separated embeddings of image-text pairs post router processing. Moreover, we also investigate the effectiveness of our router when trained with different numbers of modality labels. Results in Figure 6 demonstrate

Method	VQA-RAD		SLAKE		PathVQA		Act.	Rank
	Open	Closed	Open	Closed	Open	Closed		
LLaVA-Med (LLama7B) with LoRA	58.22 (-3.30)	82.13 (-2.06)	81.29 (-1.79)	83.37 (-1.97)	34.33 (-3.62)	90.12 (-1.09)	7B	128
LLaVA-Med (Vicuna7B) with LoRA	61.37 (-3.02)	80.03 (-1.95)	82.02 (-2.69)	81.74 (-1.43)	36.78 (-2.09)	90.67 (-0.98)	7B	128
Med-LoRAMoE (Phi2)	58.12 (**-0.43**)	82.35 (**-0.37**)	83.58 (**-0.12**)	84.85 (_-0.49_)	32.62 (**-0.63**)	91.18 (**-0.80**)	3.6B	256
Med-LoRAMoE (Phi2)	57.20 (_-1.35_)	81.75 (_-0.97_)	83.95 (_-0.35_)	84.37 (-1.21)	33.03 (_-0.98_)	90.83 (-1.15)	3.6B	128
Med-LoRAMoE (StableLM)	47.83 (-2.25)	79.04 (-1.03)	82.12 (-1.04)	82.45 (-0.96)	33.28 (-1.46)	90.80 (_-0.89_)	2.0B	256
Med-LoRAMoE (StableLM)	45.74 (-2.65)	78.31 (-1.76)	82.27 (-0.89)	83.17 (**-0.24**)	32.20 (-2.53)	90.62 (-1.08)	2.0B	128

Table 7: Comparison of models with LoRA across VQA in open and closed settings. Deltas indicate performance changes compared to models without LoRA. The smallest changes are in bold while the second smallest are underlined.

that the training of our router only require a small set of modality labels without incurring much computational cost.

Method	MLP Parameters	VQA-RAD		SLAKE		PathVQA	
		Open	Closed	Open	Closed	Open	Closed
a	0.02MB(MLP x 1)	44.53	76.48	81.85	81.78	31.05	90.42
b	0.02MB(MLP x 1)	**45.74**	**78.31**	82.27	**83.17**	**32.20**	90.62
c	1.00MB(MLP x 2)	45.03	77.64	**82.36**	82.98	32.13	**90.67**
d	1.13MB(MLP x 3)	44.98	77.85	81.77	82.09	31.87	90.17
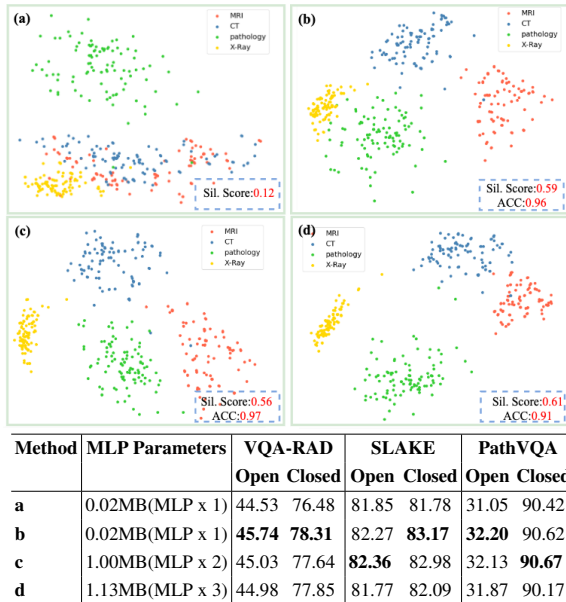
Figure 3: Visualization of task embeddings and performance using routers under varied settings. Silhouette score (sil. score) denotes superior task differentiation. Supplementary Figure 10 illustrates Phi2's embeddings.

Image and Text Specialization in Experts: As shown in Figure 4, we visualize the domain specificity of experts when processing MRI inputs. Each expert shows distinct preferences for handling text or image information. For example, Expert 1 mainly handles text data, while Expert 2 has no preference for text or image data. Expert 3 focuses on image data, whereas Expert 4 specializes in text data. This differentiation highlights the Router's ability to enhance MoE model efficiency and performance by assigning tasks to suitable experts.

Domain Specialization of Images in Experts: Figure 5 visualizes the activation states of MoE experts during inference. We sample 200 data points

across different modalities and datasets, identifying the top-1 expert with the most activations in each MoE layer. The visualization shows domain specialization for different input image modalities: Expert 1 and Expert 2 for CT, Expert 2 and Expert 3 for MRI, Expert 4 for Pathology, and Expert 1 for X-Ray. This specialization, due to our router and meta experts, enhances MoE performance by encouraging each expert to focus on specific modalities and collaborate with the meta expert for global information. However, visualizing expert activations for four modalities handled by traditional routers in each MoE layer reveals fused patterns, resulting in weaker interpretability and performance.

Cost Analysis: Table 8 illustrates the cost efficiency of our models compared to LLaVA-Med. Particularly, Med-LoRAMoE models show significant reductions in training GPU memory usage and inference time. For example, Med-LoRAMoE (StableLM) requires only 8GB of GPU memory and 3 seconds for inference, demonstrating high efficiency for deployment, and making our approach more appealing to practical resource-constrained clinical settings.

Model	Mod. Size	Tra. GPU	Inf. Time	Load Mem.
LLaVA-Med	7B	>24GB	5s	20GB
Med-MoE (Phi2)	3.6B	23GB	3s	13.4GB
Med-MoE (StableLM)	2.0B	12.5GB	3s	10.5GB
Med-LoRAMoE (Phi2)	3.6B	13.5GB	3s	13.7GB
Med-LoRAMoE (StableLM)	2.0B	8GB	3s	10.8GB

Table 8: Cost efficiency comparison of different models.

Effect of the Number of Activated Experts: The router's top-k selection is typically 1 or 2 in many existing MoE works, as more would negate sparsity benefits and increase memory overhead. We evaluate performance with different numbers of activated experts, as shown in Table 9. Ultimately, we chose 4 experts with top-2 activations for balanced performance and overhead.

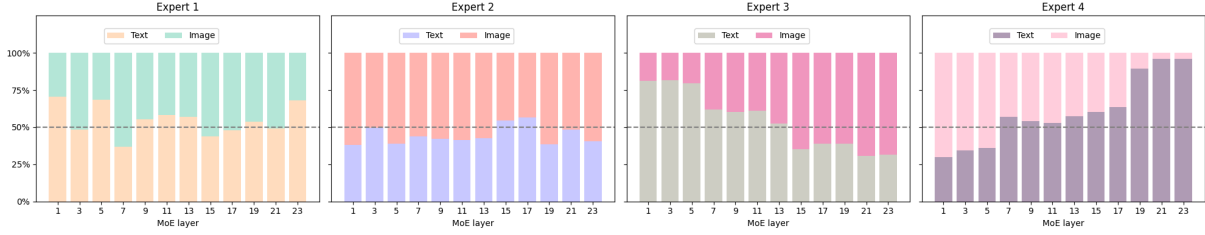**Effect of the Number of Experts:** In MoE ap-

Figure 4: Visualization of expert specialization in processing image and text tokens under the MRI modality. Results for other modalities are in Supplementary Figure 11.
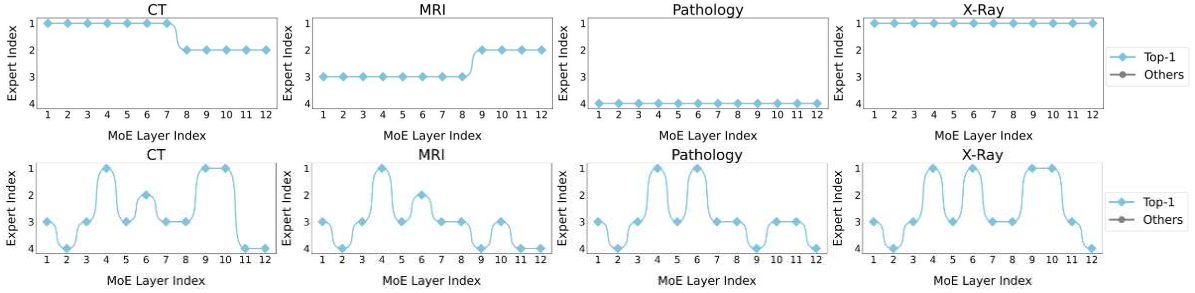

Figure 5: **Upper:** Expert activations for four modalities handled by our router of Med-MoE in each MoE layer. **Lower:** Expert activations for four modalities handled by the standard learned router in each MoE layer.

Top-k (Experts=4)	1 Expert	2 Experts
VQA-RAD (Open)	46.7	47.2 (**+0.5**)
VQA-RAD (Closed)	83.4	83.8 (**+0.4**)
SLAKE (Open)	82.1	82.3 (**+0.2**)
SLAKE (Closed)	83.2	84.9 (**+1.7**)
PathVQA (Open)	33.9	34.1 (**+0.2**)
PathVQA (Closed)	90.9	91.8 (**+0.9**)
Time	7h	8h

Table 9: Performance with varying activated experts

Experts (Top-k=2)	2 Experts	4 Experts(Ours)	6 Experts
VQA-RAD (Open)	47.03 (**-0.8**)	47.83	48.03 (**+0.2**)
VQA-RAD (Closed)	77.54 (**-1.5**)	79.04	78.84 (**-0.2**)
SLAKE (Open)	81.52 (**-0.6**)	82.12	82.42 (**+0.3**)
SLAKE (Closed)	81.45 (**-1.0**)	82.45	82.65 (**+0.2**)
PathVQA (Open)	32.78 (**-0.5**)	33.28	33.58 (**+0.3**)
PathVQA (Closed)	90.60 (**-0.2**)	90.80	91.10 (**+0.3**)
Time	6h	8h	11h

Table 10: Performance with varied expert number

plications, choosing the right number of experts is crucial for balancing performance and computational efficiency. We evaluate configurations with 2, 4, and 6 experts, each with the top-2 activations. Results in Table 10 show that using 4 experts achieves the best balance between performance and cost efficiency. While increasing to 6 experts offers slight performance gains, it significantly increases computational time and memory usage. On the

other hand, reducing the number to 2 experts leads to decreased performance across all tasks.

Model	PneumoniaMNIST	OrganCMnist	DermaMNIST	BloodMNIST
AutoKeras	87.8	87.9	74.9	96.1
BiomedGPT	90.8	88.9	72.3	97.2
Med-MoE-Phi2	91.4	89.9	-	-
Med-MoE-StableLM	89.3	88.6	-	-
Med-MoE (Scaling)	91.6 (+2.3)	90.1 (+1.5)	80.4 (+5.7)	95.1

Table 11: Performance after training data scaling

Effect of Scaling Training Data: To validate the performance of our model with a larger training dataset, we increase the classification data from BloodMNIST and DermaMNIST (Yang et al., 2023, 2021) to a total of 36.7K for the classification task in Stage 3. After training with Med-MoE (StableLM), as shown in Table 11, we observed a significant improvement in classification accuracy. The addition of new data clearly demonstrates a substantial enhancement in our model's capabilities, indicating that our approach is adaptable to more modalities and larger datasets, showcasing potential for practical applications.

5   Related Work

Medical MLLMs Advancements in Medical MLLMs, such as Med-Flamingo (Moor et al., 2023), Med-PaLM M (Singhal et al., 2023), and LLaVA-Med (Li et al., 2024a), have significantly impacted medical diagnostics and patient care,

building on general AI models like ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023). These models enhance few-shot learning, medical question answering, and conversational AI, demonstrating the potential of specialized MLLMs in healthcare. Biomedical chatbots like ChatDoctor (Yunxiang et al., 2023) and Visual Med-Alpaca highlight the benefits of domain-specific fine-tuning. However, their application in resource-constrained hospital settings remains underexplored, emphasizing the need for cost-efficient MLLMs in clinical contexts.

MoE in MLLMs MoE in MLLMs addresses task conflicts in multi-task learning and offers a cost-efficient scaling method. The first approach (Xu et al., 2024a; Chen et al., 2024; Gou et al., 2023) uses Top-1 activation to assign different tasks to different experts, avoiding performance degradation from task data conflicts but overlooking modal biases within the same task type. The second approach (Lin et al., 2024; Li et al., 2024b; Lee et al., 2024; Liu et al., 2024b; Dai et al., 2024) replaces FFN layers in LLMs with MoE structures using multiple expert activations, achieving improvements with minimal additional parameters. However, visualizing the expert activations in routers shows these methods often fail to specialize experts effectively, limiting interpretability and performance with diverse data modalities (Fan et al., 2024). A specialized MoE architecture tailored to the medical domain is needed to leverage modality-specific information and improve performance with a smaller LLM backbone.

6 Conclusion

We have introduced Med-MoE, a lightweight framework for multimodal medical tasks, addressing both discriminative and generative needs. Optimized for resource-constrained environments, Med-MoE involves aligning medical images with language model tokens, task-specific instruction tuning, and domain-specific expert fine-tuning. Our approach reduces activated parameters while maintaining or surpassing state-of-the-art performance. Our experiments on VQA-RAD, SLAKE, and Path-VQA validate Med-MoE's effectiveness and efficiency. This model offers a practical solution for deploying advanced medical AI in diverse and resource-limited clinical settings.

7 Discussion and Limitations

Our work primarily sought to develop a smaller, more cost-efficient Multimodal Large Language Model (MLLM) for the medical field, diverging from the current trend focused on creating larger and more robust models. We posit that in practical applications, especially in resource-constrained environments like mobile devices, smaller models could be more advantageous. This approach not only addresses the practical limitations of deploying large-scale models in routine clinical settings but also explores the feasibility of using leaner models without compromising on performance, fostering broader accessibility and application.

However, our approach faces several limitations. First, there is a notable scarcity of training data in the medical domain, largely due to the sensitivity and privacy concerns associated with medical data. Generating synthetic data through methods like those used for GPT-4V can be problematic in this context, and many datasets require labor-intensive manual annotations by medical professionals. This is both costly and limits the scalability of data generation efforts. As illustrated in Supplementary Figure 9, our model occasionally fails, particularly with more complex open-ended questions that demand precise medical knowledge.

Furthermore, the inherent requirement for medical applications to provide trustworthy explanations and confidence scores poses another challenge. Ensuring that the model outputs are not only accurate but also accompanied by reliable justifications is crucial, especially in a field where decisions have significant health implications. This necessity heightens the importance of building a trustworthy MLLM that can articulate its reasoning processes clearly and provide confidence levels, thereby enhancing the reliability and safety of AI applications in healthcare.

Acknowledgments

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. 2022. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062.

Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. 2023. Visual question answering: A survey on techniques and common trends in recent literature. *arXiv preprint arXiv:2305.11033*.

Mateusz Dubiel, Yasmine Barghouti, Kristina Kudryavtseva, and Luis A Leiva. 2024. On-device query intent prediction with lightweight llms to support ubiquitous conversations. *Scientific Reports*, 14(1):12731.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

Dongyang Fan, Bettina Messmer, and Martin Jaggi. 2024. Towards an empirical understanding of moe design choices. *arXiv preprint arXiv:2402.13089*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2023. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*.

Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*.

Jiaao He, Jiezhong Qiu, Aohan Zeng, Zhilin Yang, Jidong Zhai, and Jie Tang. 2021. Fastmoe: A fast mixture-of-expert training system. *arXiv preprint arXiv:2103.13262*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Haifeng Jin, Qingquan Song, and Xia Hu. 2019. Autokeras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1946–1956.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. 2024. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*.

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

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2022. Self-supervised vision-language pretraining for medical visual question answering. *arXiv preprint arXiv:2211.13594*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2024b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*.

Weibin Liao, Yinghao Zhu, Xinyuan Wang, Cehngwei Pan, Yasha Wang, and Liantao Ma. 2024. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023. Q2atransformer: Improving medical vqa via an answer querying decoder. *arXiv preprint arXiv:2304.01611*.

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, et al. 2024b. Mixture of insightful experts (mote): The synergy of thought chains and expert mixtures in self-alignment. *arXiv preprint arXiv:2405.00557*.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.

Yalin Miao, Shuyun He, WenFang Cheng, Guodong Li, and Meng Tong. 2022. Research on visual question answering based on dynamic memory network model of multiple attention mechanisms. *Scientific Reports*, 12(1):16758.

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

OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774. *Preprint*, arXiv:2303.08774.

Lena Petersson, Ingrid Larsson, Jens M Nygren, Per Nilsen, Margit Neher, Julie E Reed, Daniel Tyskbo, and Petra Svedberg. 2022. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in sweden. *BMC Health Services Research*, 22(1):850.

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

Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*.

Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Jingwei Xu, Junyu Lai, and Yunpeng Huang. 2024a. Meteora: Multiple-tasks embedded lora for large language models. *arXiv preprint arXiv:2405.13053*.

Xi Xu, Jianqiang Li, Zhichao Zhu, Linna Zhao, Huina Wang, Changwei Song, Yining Chen, Qing Zhao, Jijiang Yang, and Yan Pei. 2024b. A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. *Bioengineering*, 11(3):219.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2024. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36.

Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 191–195. IEEE.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Yubiao Yue and Zhenzhang Li. 2024. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

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

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

A   Appendix

Calculation Formulas for Open Setting Metrics

1. Recall

Recall is calculated using the following formula:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

where TP (true positives) is the number of words in both the candidate and the reference, and FN (false negatives) is the number of words in the reference but not in the candidate.

2. Exact Match Score

Exact Match Score is calculated using the following formula:

$$\text{EMS} = \frac{\text{Number of matching words}}{\text{Total number of candidate words}} \tag{7}$$

This formula calculates the ratio of the number of matching words in the candidate and the reference to the total number of words in the candidate.

3. BLEU Score

The BLEU Score is calculated using the following steps:

- Calculate the modified precision p_n for each n-gram up to n:

$$p_n = \frac{\sum_{C \in \text{Candidates}} \sum_{ng \in C} \min(\text{Count}(ng), \text{Count}_{\max}(ng))}{\sum_{C \in \text{Candidates}} \sum_{ng \in C} \text{Count}(ng)}$$
(8)

where ng is the n-gram, $\text{Count}(ng)$ is its count in the candidate, and $\text{Count}_{\max}(ng)$ is its maximum count in the reference.

- Calculate the brevity penalty (BP):

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1 - \frac{r}{c})} & \text{if } c \leq r \end{cases}$$
(9)

where c is the length of the candidate sentence and r is the length of the reference sentence.

- Combine the modified precision and brevity penalty to compute the BLEU score:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{i=1}^{n} w_i \log p_i\right)$$
(10)

where w_i are the weights assigned to each n-gram precision.

A.1 Dataset information

VQA-RAD (Lau et al., 2018) contains 3,515 QA pairs and 315 radiology images, with questions covering 11 categories and a mix of closed-ended and open-ended types. SLAKE (Liu et al., 2021) comprises 642 radiology images and over 7,000 QA pairs, including segmentation masks and object detection bounding boxes. PathVQA (He et al., 2020) includes 4,998 pathology images with 32,799 QA pairs, focusing on aspects like location, shape, color, and appearance, categorized into open-ended and closed-ended types. The Pneumonia-MNIST dataset focuses on pediatric chest radiographs for binary classification of pneumonia versus normal, using 4,708 training and 624 test images. OrganCMNIST (Yang et al., 2023) classifies 11 human body organs with 12,975 training and 8,216 testing images. To ensure a fair comparison with LLaVA-Med, we did not use the additional image classification training datasets when evaluating VQA. For evaluation, we use test sets from these widely recognized medical VQA datasets and additionally assess classification performance.
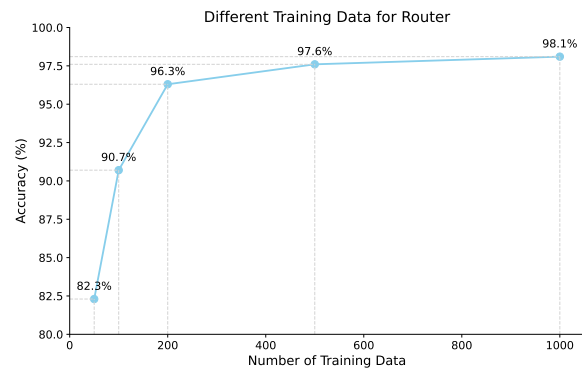

Figure 6: Performance of router predictions with different domain-labeled training data

Config	Stage I	Stage II	Stage III
Deepspeed	Zero2, Zero2, Zero2 offload		
Image encoder	CLIP-Large		
Feature select layer	-2		
Image projector	2 Linear layers with GeLU		
Epoch (same as LLaVA-Med)	1	3	9
Learning rate	1e-3	2e-5	2e-5
Learning rate schedule	Cosine		
Weight decay	0.0		
Text max length	2048		
Batch size per GPU	2		
GPU	8 × 3090-24G		
Precision	Bf16		
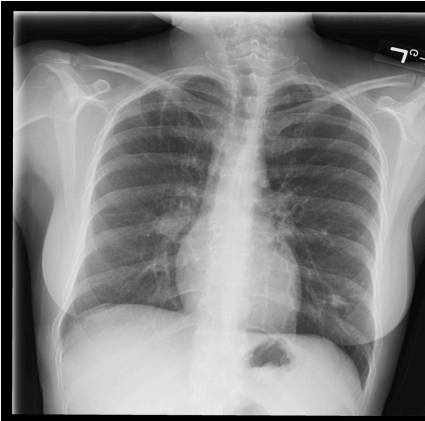
Table 12: Our experimental hyperparameters

Stage	Data Source	Sample Size
Stage 1	llava_med_alignment_500k.json	500K
Stage 2	instruct_60k_inline_mention	60K
Stage 3	**VQA:** RAD-VQA, SLAKE, Path-VQA: 27K	44K
	Classification: PneumoniaMNIST, OrganCMNIST: 17K	

Table 13: Summary of Data Utilized Across Training Stages

Name	Experts	Activated Experts	MoE Layers	Embedding	Width	Layers	FFN	FFN Factor	Heads	Activated Param	Total Param
StableLM-1.6B	-	-	-	100352	2560	32	10240	2	32	1.6B	1.6B
Med-MoE (StableLM-4x1.6B)	4	2	16	100352	2560	32	10240	2	32	2.0B	2.9B
Phi2-2.7B	-	-	-	51200	2560	32	10240	2	32	2.7B	2.7B
Med-MoE (Phi2-4x2.7B)	4	2	16	51200	2560	32	10240	2	32	3.6B	5.3B

Table 14: Comparison of different models in terms of various parameters.

GPT4 Evaluation

Assuming you are a doctor, please rate the following three answers about this medical picture from 0 to 10, and tell me why.

From the perspective of a medical professional:

1.This explanation is detailed and accurate, so I'd give it an 8 out of 10. It mentions the specific posteroanterior view, which is essential for proper interpretation of a chest X-ray. It also lists the structures that are typically assessed with this imaging technique, though it does not provide much detail about what might be diagnosed with this information.

2.I would rate this explanation a 7 out of 10. While it correctly identifies the type of chest X-ray, it does not offer any additional details beyond what the first explanation provides. The term "surrounding structures" is somewhat vague and could be expanded upon for clarity.

3.This response gets a 9 out of 10. It offers an extensive overview of the anatomy visible on a chest X-ray, noting the appearance of normal lung fields, the positioning of the heart, and the visibility of the diaphragm. It goes further to explain the clinical utility of a chest X-ray, detailing various conditions it can help diagnose. This level of detail is very informative and has significant educational value for understanding both the anatomy and the clinical reasoning behind the use of chest X-rays.

The ratings are based on the depth and utility of the information provided. The third response is rated highest because it provides a thorough explanation that could enhance understanding in a clinical education context and assist in patient care decisions.

Q:What is shown in this image?

MoE-TinyMed: The image is a posteroanterior chest X-ray, which is a common diagnostic imaging technique used to visualize the structures within the chest, including the lungs, heart, and bones of the chest and spine.

LLaVA-Med: The image is a postero-anterior chest X-ray, which shows the patient's chest, including the heart, lungs, and surrounding structures.

GPT4V: This image is a chest X-ray, a type of radiograph that is used in medical settings to examine the chest cavity. It shows the bones of the shoulder girdle and the upper arms, as well as the bony thorax which includes the ribs, the thoracic portion of the spine, and the sternum. It also shows the lung fields, which should normally appear darker due to the air content, the heart, which is usually visible in the center as a denser area, and the diaphragm separating the chest cavity from the abdominal cavity.

Figure 7: An example showcasing our method's ability to answer medical imaging questions with performance nearing or even surpassing that of LLaVA-Med under GPT-4V evaluation.
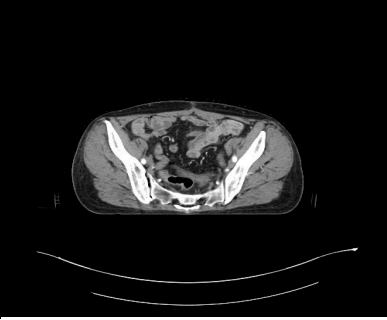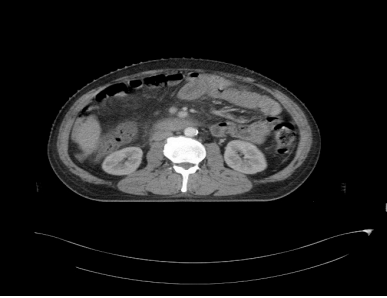
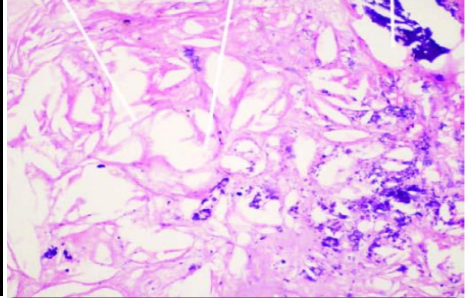| Case from RAD-VQA	Case from SLAKE	Case from Path-VQA
Q:The image is taken in what plane? A:Axial	Q:Which is smaller in this image, colon or small bowel? A:Colon	Q:Does typical tuberculous exudate show obvious lesion? A:No
Q:Is there air outside the bowel walls? A:No	Q:What is the shape of spinal cord in this image? A:Circular	Q:Is there narrowing of the lumen of coronary due to fully developed atheromatous plaque which has dystrophic calcification in its core? A:No

Figure 8: More Medical VQA cases from VQA-RAD, SLAKE, and Path-VQA. our Med-MoE generates expected responses for medical image queries.
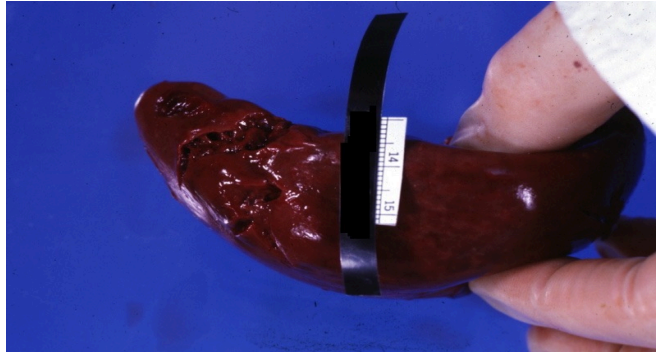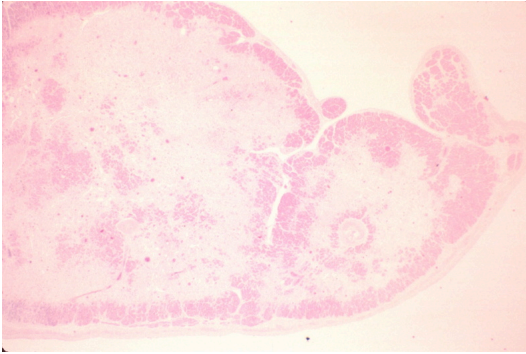
Q: What is present?
A:endocrine
GT:cardiovascular

Q: What done external view of lacerations of capsule done during?
A:dissection
GT:done surgical procedure
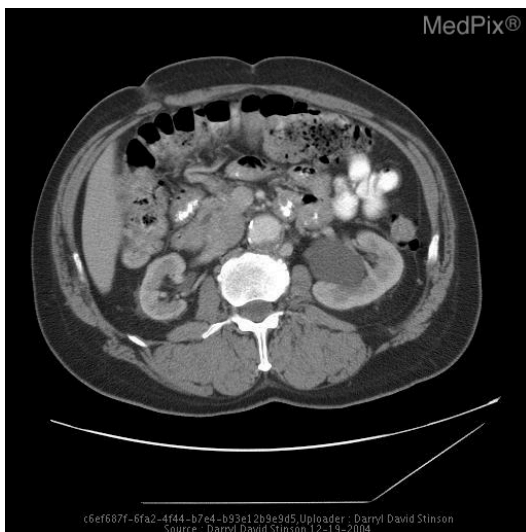
Q: What are the hyperdense lesions noted at the edges of the aorta?
A:Calcifications
GT:Calcified atherosclerosis

Q: What structures are involved?
A:basal ganglia, cerebellum, cerebral cortex
GT:Caudate, putamen, left parietal

Figure 9: Incorrect cases: Med-VQA examples in OPEN setting requiring precise and specialized medical knowledge.

Silhouette Score:0.173

Silhouette Score:0.579
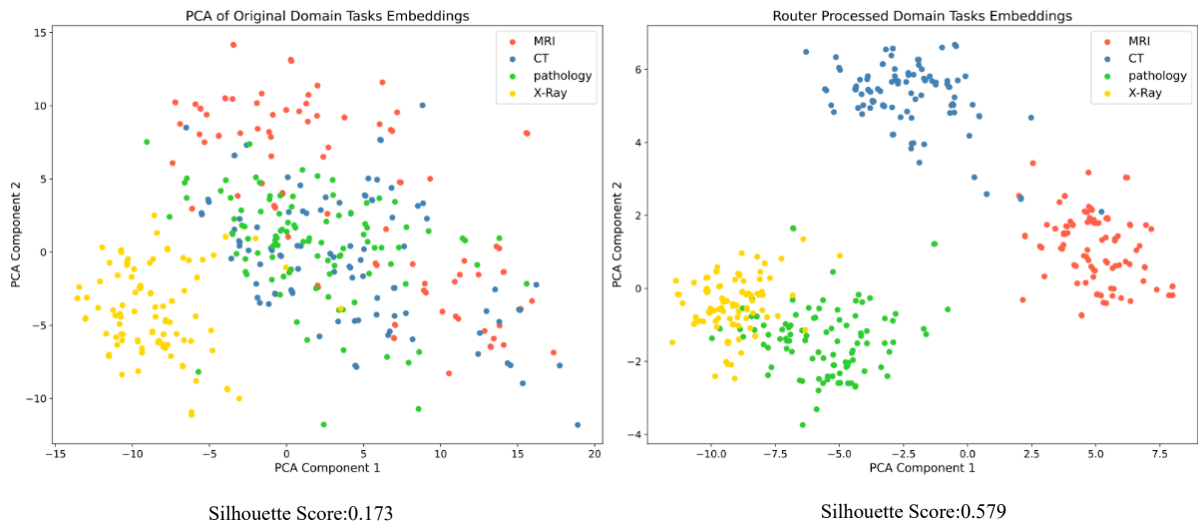
(a). Embedding visualization for domain-specific tasks using Router in MoE-TinyMed-Phi2.

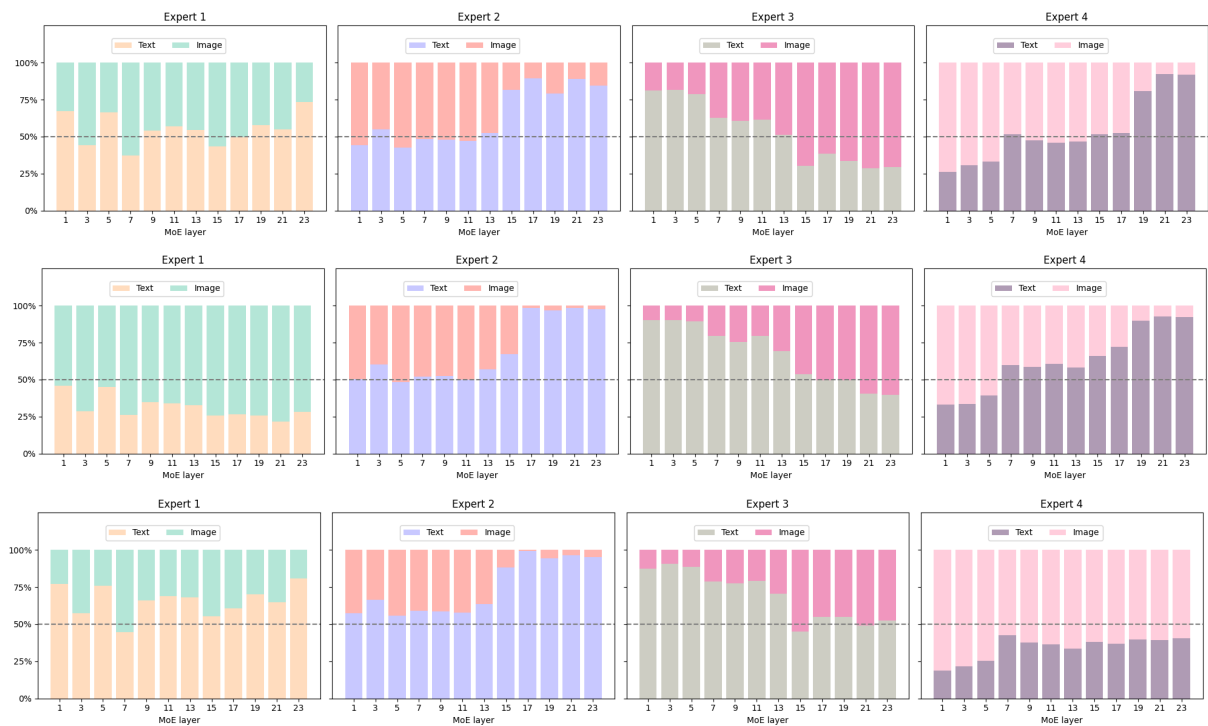Figure 10: Experts Routing Visualization On Phi2

Figure 11: The activation proportions for text and image processing in other modalities