Uni-Med: A Unified Medical Generalist Foundation Model For Multi-Task Learning Via Connector-MoE

Xun Zhu[1]   Ying Hu[1]   Fanbin Mo[2]   Miao Li[1, ✉]   Ji Wu[1, 3, 4]

[1] Department of Electronic Engineering, Tsinghua University
[2] Beijing University of Posts and Telecommunications   [3] College of AI, Tsinghua University
[4] Beijing National Research Center for Information Science and Technology
{zhu-x24, yinghu_yh}@mails.tsinghua.edu.cn   mofanbin@bupt.edu.cn
{miao-li, wuji_ee}@tsinghua.edu.cn   ✉ corresponding author

Abstract

Multi-modal large language models (MLLMs) have shown impressive capabilities as a general-purpose interface for various visual and linguistic tasks. However, building a unified MLLM for multi-task learning in the medical field remains a thorny challenge. To mitigate the tug-of-war problem of multi-modal multi-task optimization in MLLMs, recent advances primarily focus on improving the LLM components, while neglecting the connector that bridges the gap between modalities. In this paper, we introduce Uni-Med, a novel medical generalist foundation model which consists of a universal visual feature extraction module, a connector mixture-of-experts (CMoE) module, and an LLM. Benefiting from the proposed CMoE that leverages a well-designed router with a mixture of projection experts at the connector, Uni-Med achieves efficient solution to the tug-of-war problem and can perform six different medical tasks including question answering, visual question answering, report generation, referring expression comprehension, referring expression generation and image classification. To the best of our knowledge, Uni-Med is the first effort to tackle multi-task interference at the connector in MLLMs. Extensive ablation experiments validate the effectiveness of introducing CMoE under any configuration, with up to an average 8% performance gains. We further provide interpretation analysis of the tug-of-war problem from the perspective of gradient optimization and parameter statistics. Compared to previous state-of-the-art medical MLLMs, Uni-Med achieves competitive or superior evaluation metrics on diverse tasks. Code and resources are available at https://github.com/MSIIP/Uni-Med.

1 Introduction

Driven by the growth of datasets, the increase in model size, and advances in generative language foundation models [Achiam *et al.*, 2023; Touvron *et al.*, 2023], multi-modal large language models (MLLMs) now offer unprecedented abilities as general-purpose interfaces. These advancements are spurring innovation across various visual and linguistic tasks [Chen *et al.*, 2023a; Lyu *et al.*, 2023; Su *et al.*, 2023]. While significant strides have been made in building a unified foundation model for natural scenery [Chen *et al.*, 2022; Lu *et al.*, 2022, 2023], the development of generalist medical artificial intelligence is still in its early stages [Moor *et al.*, 2023a].

The goal of a unified and generalist medical foundation model is to enable joint training on massive medical datasets. This model aims to handle multiple tasks and modalities within a single architecture with shared parameters [Zhang *et al.*, 2023; Li *et al.*, 2024]. It seeks to eliminate the need for task-specific modules and further fine-tuning, thereby revolutionizing the traditional task-specific
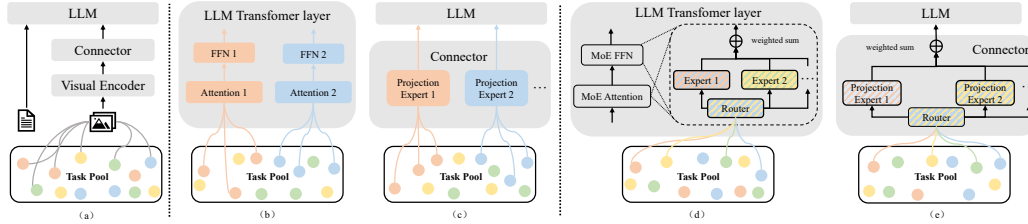
Figure 1: Three hypotheses and corresponding architectural implementations for multi-task learning in MLLMs. (a) Synergy hypothesis. (b)-(c) Conflict hypothesis in LLM and connector, respectively. (d)-(e) Conflict-synergy coexist hypothesis in LLM and connector, respectively.

approach to model development [Wu *et al.*, 2023b; Tu *et al.*, 2024]. However, existing open-source efforts have not yet fully achieved these ambitious goals.

A key challenge in creating a unified medical foundation model is the complexity of multi-modal, multi-task learning, often exacerbated by the tug-of-war problem [Hadsell *et al.*, 2020]. Inherent task conflicts and data imbalances can cause interference during the simultaneous learning of different tasks. This problem is particularly acute in the medical field, where tasks and modalities are highly specialized and diverse. As a result, the performance of each task may degrade compared to task-specialized models [Yu *et al.*, 2020; Zhu *et al.*, 2022].

To mitigate the tug-of-war problem in multi-task learning, recent advances introduce the well-known Mixture-of-Experts (MoE) [Jacobs *et al.*, 1991] into MLLMs. Figure 1 illustrates three distinct hypotheses and their corresponding architectural implementations for multi-task learning in MLLMs. The first "synergy hypothesis" suggests that all tasks benefit from a fully shared backbone comprising a visual encoder, connector, and language model, which is the standard architecture for MLLMs. The second "conflict hypothesis", proposes that each task requires its own specific adaptations, thereby preventing knowledge sharing among tasks. The third "conflict-synergy coexistence hypothesis", posits that all tasks share multi-task adaptations, which reduces interference and promotes more efficient knowledge sharing. However, current research [Zadouri *et al.*, 2023; Gou *et al.*, 2023; Liu *et al.*, 2023b; Lin *et al.*, 2024] mainly tailors the MoE approach to the language model components, overlooking the potential benefits of exploring and enhancing the connector in MLLMs. Furthermore, the optimization of the tug-of-war problem lacks a detailed, explainable analysis.

In this study, we first identify a tug-of-war problem in multi-task learning at the connector level within standard MLLM architectures. This issue indicates that different tasks may emphasize different types of features in multi-modal, multi-task scenarios. Consequently, a fully shared connector may fall short as it cannot accommodate the diverse modal features required by each task. Drawing inspiration from the successful application of MoE in LLMs, we introduce Connector-MoE (CMoE), a novel approach that employs a mixture of projection experts to align visual and language embedding spaces effectively, thus mitigating the tug-of-war problem. As a pioneering effort in constructing a unified generalist foundation model in the medical field, we present Uni-Med. This model comprises a universal visual feature extraction module, a CMoE module, and an LMM. Uni-Med demonstrates impressive performance across six distinct medical tasks, with minimal training computational overhead. It achieves joint training on 12 datasets on a single A800 in under 10 hours. The effectiveness and generalization of CMoE are underscored through ablation experiments. Additionally, an interpretable analysis reveals that Uni-Med provides a superior solution to the tug-of-war problem at the connector level. Overall, Uni-Med delivers competitive or even superior performance compared to open-source, state-of-the-art medical MLLMs on all test sets. Our contributions can be summarized as:

- We present Uni-Med, an open-source medical generalist foundation model with a unified interface and shared parameters, which can perform six different medical tasks including question answering, visual question answering, report generation, referring expression comprehension, referring expression generation and image classification.

- We propose CMoE, a well-designed connector component for MLLMs, which significantly outperforms baselines under any configuration, with up to an average 8% performance gains. To our knowledge, Uni-Med is the first attempt to focus on the connector in MLLMs to mitigate the tug-of-war problem, which is critical but has always been overlooked.

- Focusing on the question of how the tug-of-war problem is optimized, which has never been quantitatively discussed, we provide detailed interpretability analysis and instructive findings from the perspective of gradient optimization and parameter statistics.
- Uni-Med achieves competitive or superior performance compared to the open-source, state-of-the-art medical MLLMs on test set of diverse tasks and datasets, which demonstrates the huge potential of medical generalist foundation models.

2 Related work

Medical foundation models   The increasing availability of medical data, as well as advances in multi-modal LLM technologies, have paved the way for the emergence of medical foundational models. Med-Flamingo [Moor *et al.*, 2023b] continues pre-training on paired and interleaved medical image-text data based on OpenFlamingo [Awadalla *et al.*, 2023]. LLaVA-Med [Li *et al.*, 2024] curates a medical multi-modal instruction following dataset and fine-tunes LLaVA [Liu *et al.*, 2024a] with it. XrayGPT [Thawkar *et al.*, 2023] can analyze and answer open-ended questions about chest X-rays. BiomedGPT [Zhang *et al.*, 2023] is a multi-task foundation model pretrained on a diverse source of medical images, literature, and clinical notes. However, most of these efforts require further fine-tuning on task-specific data to support downstream applications. One step further, the generalist foundation model uses the same weight to excel at various tasks without fine-tuning. RadFM [Wu *et al.*, 2023b] is dedicated to build a generalist foundation model for radiology. Med-PaLM M [Tu *et al.*, 2024] is directly trained in a unified framework to jointly handle many tasks, which is perhaps most similar to our effort, but it does not provide access for usage. In addition, recent studies [Wu *et al.*, 2023a; Yan *et al.*, 2024; Xia *et al.*, 2024] have suggested the the necessity for a more comprehensive and detailed evaluation of the capabilities of medical MLLMs.

MoE in multi-task learning   MoE is originally considered to increase the model capacity [Riquelme *et al.*, 2021; Fedus *et al.*, 2022] and gains popularity in mitigating multi-task interference [Chen *et al.*, 2023e, 2024]. It achieves this by utilizing a router to determine the token set handled by each expert, thus reducing interference between different types of samples. Recent studies have focused on combining MoE with LLM, such as MoE-LLaVA [Lin *et al.*, 2024] and Mixtral 8x7B [Jiang *et al.*, 2024], or combining MoE with one of the representative parameter-efficient tuning techniques, i.e., LoRA [Hu *et al.*, 2021], such as Octavius [Chen *et al.*, 2023d], MoCLE [Gou *et al.*, 2023], MTLoRA [Agiza *et al.*, 2024] and MOELoRA[Liu *et al.*, 2024b]. However, neither of them introduces MoE into the connector component for MLLMs. Furthermore, there is a lack of clear and explicit interpretable analysis on how the multi-task interference is mitigated through the use of MoE.

Cross-modality connector in MLLM   The connector between the multi-modal encoder and the LLM is critical in aligning multi-modal features [Song *et al.*, 2023]. One of the most popular paradigms is to map multi-modal features into a feature space that aligns with language, such as linear projection [Liu *et al.*, 2024a] and MLP projection [Liu *et al.*, 2023a; Chen *et al.*, 2023c]. Another paradigm is to transform multi-modal features into multi-modal tokens that are consistent with the embedded representation space of LLM, such as cross-attention [Li *et al.*, 2022; Ye *et al.*, 2023b, 2024], perceiver resampler [Alayrac *et al.*, 2022; Peng *et al.*, 2023] and Q-Former [Li *et al.*, 2023; Zhu *et al.*, 2023]. However, existing paradigms use the same connector when processing the same modal data for different tasks, ignoring the imperative to acquire distinct alignment patterns tailored to the demands of each task.

3 Methodology

3.1 Preliminaries

3.1.1 Multi-task interference

To quantify the intricate tug-of-war problem in a unified foundation model, we provide interpretability from the perspective of gradient optimization and parameter statistics.

Perspective of gradient optimization   When optimizing the shared parameters θ according to task j, the change in the update direction of loss L_i for task i can be defined as [Zhu *et al.*, 2022]:
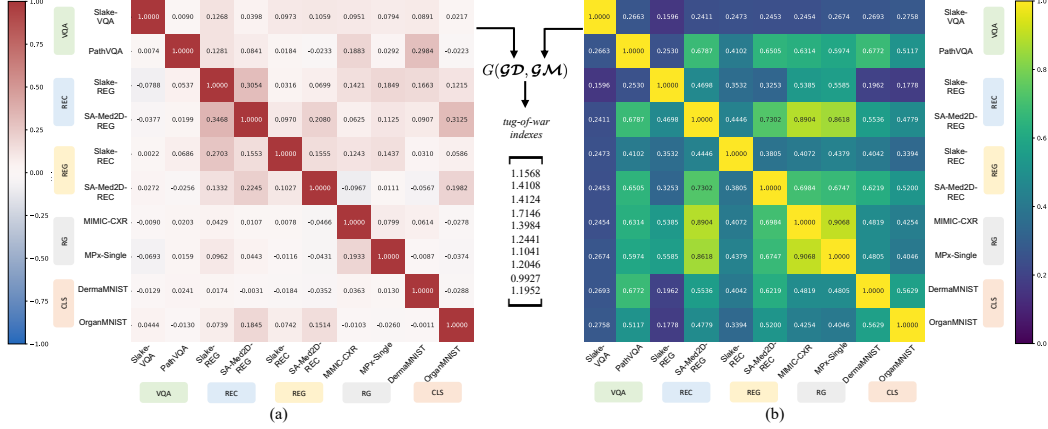
Figure 2: Dataset-level multi-task interference of the synergy hypothesis model at the connector in MLLMs. (a) Perspective of gradient direction $\mathcal{GD}$. (b) Perspective of gradient magnitude $\mathcal{GM}$.

$$\Delta_j L_i\left(x_i\right) \doteq \mathbb{E}_{x_j}\left(L_i\left(x_i; \theta\right) - L_i\left(x_i; \theta - \lambda \frac{\nabla_\theta L_j\left(x_j\right)}{\left\|\nabla_\theta L_j\left(x_j\right)\right\|_2}\right)\right) \approx \lambda \mathbb{E}_{x_j}\left(\frac{\nabla_\theta L_j\left(x_j\right)^T}{\left\|\nabla_\theta L_j\left(x_j\right)\right\|_2} \nabla_\theta L_i\left(x_i\right)\right)$$
(1)

where x_i and x_j are the sampled training batches of task i and j, respectively. The interference of task j on task i in the update direction can be quantified as:

$$\mathcal{GD}_{i,j} = \mathbb{E}_{x_i}\left(\frac{\Delta_j L_i\left(x_i\right)}{\Delta_i L_i\left(x_i\right)}\right)$$
(2)

The gradient magnitude similarity between task i and task j can be defined as:

$$\mathcal{GM}_{i,j} = \mathcal{GM}_{j,i} = \frac{2\mathbb{E}_{x_i}\left(\left\|\nabla_\theta L_i\left(x_i\right)\right\|_2\right) \mathbb{E}_{x_j}\left(\left\|\nabla_\theta L_j\left(x_j\right)\right\|_2\right)}{\left(\mathbb{E}_{x_i}\left(\left\|\nabla_\theta L_i\left(x_i\right)\right\|_2\right)\right)^2 + \left(\mathbb{E}_{x_j}\left(\left\|\nabla_\theta L_j\left(x_j\right)\right\|_2\right)\right)^2}$$
(3)

$\mathcal{GM}_{i,j}$ goes to zero when the difference in gradient magnitudes is large, indicating that some task is dominant [Yu *et al.*, 2020]. For all T tasks, we can get $\mathcal{GD}, \mathcal{GM} \in \mathbf{R}^{T \times T}$. Then, we define the tug-of-war indexes for each task in multi-task learning through the function G as follows:

$$\textit{tug-of-war indexes} = G(\mathcal{GD}, \mathcal{GM}) = \left[\sum_{j=1}^{T} \mathcal{GD}_{i,j} \cdot \mathcal{GM}_{i,j}\right]_{i=1}^{T}$$
(4)

Perspective of parameter statistics Inspired by the Gradient Positive Sign Purity proposed by Chen *et al.* [2020], we define the statistics score of a single parameter in multi-task learning:

$$\textit{statistics score} = \left|\frac{\sum_i^T \nabla_\theta L_i}{\sum_i^T \left|\nabla_\theta L_i\right|}\right|$$
(5)

where $\nabla_\theta L_i$ is the gradient for task i. The range of the statistics score is [0, 1], and a value close to 1 indicates that this parameter suffers less gradient conflict during multi-task training. Upon collecting the statistics scores of all parameters, we can intuitively demonstrate and analyze the phenomenon of multi-task interference.

To be specific, we sample 100 batches for each datasets and record the gradients to calculate all of the above metrics. Figure 2 shows the dataset-level (more granular than task-level) multi-task interference of the synergy hypothesis model at the connector in the standard MLLM architecture.

3.1.2 Mixture-of-Experts

A Mixture-of-Experts (MoE) contains a set of expert networks $E_1, E_2, ..., E_N$ along with a routing network R. For each token x_i in the input sequence $\boldsymbol{X} = \{x_i\}_{i=1}^{L}$, the output of MoE is the weighted sum of outputs from each expert, where the weight is calculated by the router:

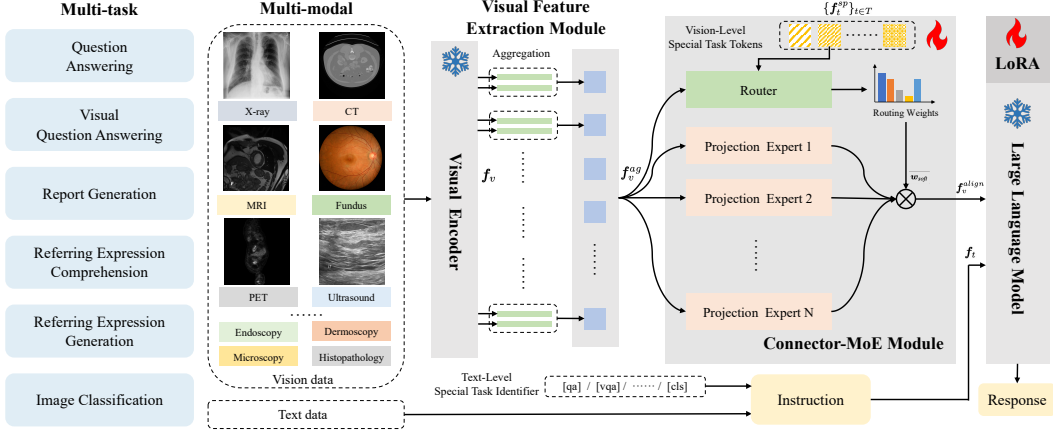$$y_i = \sum_{k=1}^{N} R(x_i)_k \cdot E_k(x_i)$$
(6)

4

Figure 3: Overall architecture of Uni-Med, which consists of a universal vision feature extraction module, a connector-MoE module and an LLM. Uni-Med can perform six different medical tasks including question answering, visual question answering, report generation, referring expression comprehension, referring expression generation and image classification.

The types of R can mainly be divided into: 1) Constant router, which assigns equal weight to each expert. 2) Hard router, which enforces one-to-one mapping between tasks and experts. 3) Sparse router, which selects Top-K experts with the maximum routing weight. 4) Soft router, which calculates the routing weights for each expert. For more details on the routing networks, see Appendix A.1.

3.2 Model Architecture

With the primary goal of achieving a unified medical generalist foundation model and mitigating the tug-of-war problem of multi-task learning in mind, we design the overall architecture of Uni-Med as illustrated in Figure 3, which contains three components: a universal vision feature extraction module, a connector-MoE module and an LLM. Detailed descriptions are presented in the following sections.

3.2.1 Visual feature extraction module

Taking one of the multi-modal medical images $\boldsymbol{I} \in \mathbf{R}^{H \times W \times C}$ as input, the visual encoder $\boldsymbol{V}_{en}$ extracts the image tokens $\boldsymbol{f}_v \in \mathbf{R}^{N_v \times D_v}$ for image perception, where $N_v = HW/P^2$ is the number of image patches and D_v is the hidden size of visual embeddings.

To alleviate the efficiency issues caused by prolonged visual input tokens during the training and inference, we scheme a resampler with a compression rate α for visual feature aggregation. Concretely, α adjacent visual tokens are concatenated and projected into one single embedding. Thus we obtain aggregated image tokens $\boldsymbol{f}_v^{ag} \in \mathbf{R}^{N_v/\alpha \times D_v \alpha}$ as follows:

$$\boldsymbol{f}_v^{ag} = resampler\left(\boldsymbol{V}_{en}\left(\boldsymbol{I}\right), \alpha\right) \tag{7}$$

3.2.2 Connector-MoE module

Aligning the visual space with the language embedding space of the large language model is a critical process, especially in the complex and diverse input of multi-task multi-modal medical image text pairs. Based on the conflict-synergy coexist hypothesis, we propose the Connector-MoE (CMoE) module, which aims to adaptively minimize task conflict and maximize task synergy at the connector. CMoE module has N projection experts $E_1, E_2, ..., E_N$, where each expert is a two-layer MLP, and a soft router R_{soft} to control the contribution of each expert.

According to Figure 2, we find that: (1) Gradient optimization conflict is common and consistent at the task level. (2) Even for the same task, there are significant differences in conflict and synergy at dataset-level. To alleviate the above problems, we randomly initialize vision-level special task tokens $\{\boldsymbol{f}_t^{sp}\}_{t \in T}$, where $\boldsymbol{f}_t^{sp} \in \mathbf{R}^{D_v \alpha}$ and T is the set of tasks. R_{soft} is a lightweight MLP designed to receive the concatenated inputs of $\boldsymbol{f}_v^{ag}$ (token level) and $\boldsymbol{f}_t^{sp}$ (task level), and calculate the routing weights $\boldsymbol{w}_{soft} \in \mathbf{R}^{D_v/\alpha \times N}$ of each expert for each image token, which can be formulated as:

Table 1: Text-level special task identifiers for different tasks.

Task	Question Answering	Visual Question Answering	Report Generation	Referring Expression Comprehension	Referring Expression Generation	Image Classification
Identifier	[qa]	[vqa]	[caption]	[refer]	[identify]	[cls]

$$\boldsymbol{w}_{soft}\left(\boldsymbol{f}_v^{ag}\right) = \sigma \cdot R_{soft}\left(\left[\boldsymbol{f}_v^{ag}, Repeat\left(\boldsymbol{f}_t^{sp}\right)\right]\right) \tag{8}$$

where $[,]$ denotes concatenation operation, σ is *SoftMax* function. Then we can obtain aligned visual tokens $\boldsymbol{f}_v^{align} \in \mathbf{R}^{N_v/\alpha \times D_t}$ through a weighted sum of all experts' output as follows:

$$\boldsymbol{f}_v^{align} = \sum_{k=1}^{N} \boldsymbol{w}_{soft,k} \cdot E_k(\boldsymbol{f}_v^{ag}) \tag{9}$$

where D_t is the hidden size of the language embedding space of the large language model and $\boldsymbol{w}_{soft,k}$ denotes the routing weight of the k-th projection expert. We discuss and analyze the effects of router type, router strategy, and number of experts in Section 4.2.1.

3.2.3 Large language model

Similar to the vision-level special task tokens, we assign the text-level special task identifiers for question answering (QA), visual question answering (VQA), report generation (RG), referring expression comprehension (REC), referring expression generation (REG) and image classification (CLS) as shown in Table 1, which can help reduce multi-task ambiguity [Chen *et al.*, 2023b]. The text prompt is designed as "<Img> < ImageFeature> </Img> [Task Identifier] Instruction", which merges the converted image features with the textual instructions. See details about our multi-task instruction template in Appendix C.

After word embedding, we can obtain textual tokens $\boldsymbol{f}_t \in \mathbf{R}^{N_t \times D_t}$, where N_t denotes the number of textual tokens. LLM generates the response $\boldsymbol{O} = \{O_i\}_{i=1}^{L}$ conditioned on the aligned visual tokens $\boldsymbol{f}_v^{align}$ and textual tokens $\boldsymbol{f}_t$ inputs in an autoregressive manner, which can be formulated as:

$$p\left(\mathbf{O}_t \mid \boldsymbol{f}_v^{align}, \boldsymbol{f}_t\right) = \prod_{i=1}^{L} p\left(O_i \mid \boldsymbol{f}_v^{align}, \boldsymbol{f}_t, O_{<i}\right) \tag{10}$$

where L is the length of output tokens. We use low-rank adaption (LoRA) [Hu *et al.*, 2021] for efficient LLM fine-tuning, which is applied to all the linear layers.

4 Experiments

4.1 Experiment settings

Tasks and datasets Text-only data is collected from MedQA [Jin *et al.*, 2021] and PubMedQA [Jin *et al.*, 2019] for the task of QA. Image-text pairs are collected from Path-VQA [He *et al.*, 2020] and Slake-VQA [Liu *et al.*, 2021] for the task of VQA, MIMIC-CXR [Johnson *et al.*, 2019] and MPx-Single [Wu *et al.*, 2023b] for the task of RG, MedMNIST v2 [Yang *et al.*, 2023] for the task of CLS. For tasks such as REG and REC that require representation of spatial locations, we use the bounding boxes of the format "$<X_{min}><Y_{min}><X_{max}><Y_{max}>$", which denotes the coordinates of objects. Then, we respectively process datasets Slake-VQA [Liu *et al.*, 2021] and SA-Med2D-20M [Ye *et al.*, 2023a] to get datasets Slake-REC, Slake-REG, SA-Med2D-REC, and SA-Med2D-REG. For a detailed description, processing and splitting of all datasets, see Appendix B.

Implementation details We adapt the open-sourced ViT-G/14 from EVA-CLIP [Fang *et al.*, 2023] and LLaMA2-Chat (7B) [Touvron *et al.*, 2023] as our visual backbone and LLM, respectively. During the training process, each task is assigned a sample rate that is calculated in proportion to the respective task's data volume. The visual backbone remains frozen with an input image resolution of 224*224 and the LLM is fine-tuned through LoRA [Hu *et al.*, 2021] with the rank of 8. The compression rate α=4 and the number of projection experts N=5. Uni-Med only requires one-stage training on a NVIDIA A800-SXM4-80GB GPU, with the first 10k iterations to warm-up and a total

of 100k iterations with a batch size of 4, which lasts roughly 10 hours. The peak learning rate is set to 1e-6 and it decays to 1e-7 following the cosine strategy. We use AdamW [Loshchilov and Hutter, 2017] optimizer with β_1=0.9, β_2=0.95 and weight decay of 0.05.

Evaluation metrics For ablation studies, we report BLEU-1 for the task of VQA, REG, and RG, IoU for the task of REC, Accuracy for the task of CLS. In addition, we use $\Delta = \frac{1}{S}\sum_{i=1}^{S}(M_{m,i} - M_{b,i})/M_{b,i} \times 100\%$ to evaluate the performance gains, where $M_{m,i}$ and $M_{b,i}$ are the metrics of our model and baseline model, S can be the number of datasets or tasks. For the overall comparison between models, we report more metrics such as F1, ROUGE, METEOR, RadGraph F1 and RadCliQ [Yu $et\ al.$, 2023]. See details at Appendix D.1.

4.2 Ablation study

4.2.1 Ablation on module design

Connector design Taking the connector of a two-layer MLP as baseline setup, we first discuss the performance of different multi-task learning hypothesis. In Table 2 (a), connectors based on conflict-synergy coexist hypothesis (CMoE with sparse / soft router) show a more holistic improvement trend in multi-task learning compared to connectors based on the conflict hypothesis (CMoE with

Table 2: Experiments of ablation study. Metrics are reported on "Slake-VQA/Path-VQA", "Slake-REC/SA-Med2D-REC", "Slake-REG/SA-Med2D-REG", "MIMIC-CXR/MPx-Single", "DermaM-NIST/OrganSMNIST" for the task of VQA, REC, REG, RG, and CLS, respectively.

Connector	Router Type	Router Strategy	VQA BLEU-1	Δ (↑)	REC IoU	Δ (↑)	REG BLEU-1	Δ (↑)	RG BLEU-1	Δ (↑)	CLS Accuracy	Δ (↑)	Total Δ (↑)
(a) Connector design													
Linear	-	-	77.90 / 56.27	-1.4%	28.44 / 11.59	-23.9%	74.98 / 55.61	-2.1%	13.80 / 15.85	-11.6%	72.47 / 69.39	-5.4%	-8.9%
MLP	-	-	79.81 / 56.48		35.18 / 16.26		74.54 / 58.42		18.55 / 15.50		76.26 / 73.64		
CMoE	Constant	-	82.74 / 57.38	2.6%	33.94 / 15.49	-4.1%	73.58 / 58.51	-0.6%	23.16 / 15.88	13.7%	75.91 / 76.50	1.7%	2.7%
	Hard	-	81.85 / **59.09**	**3.6%**	30.01 / 11.59	-21.7%	70.91 / 58.04	-2.8%	22.76 / 15.79	12.3%	**81.55 / 81.18**	**8.6%**	0.0%
	Sparse	Token	80.68 / 57.02	1.0%	37.07 / 18.41	9.3%	76.86 / 60.08	3.0%	24.02 / 15.73	15.5%	73.47 / 74.93	-1.0%	5.6%
		Token	81.79 / 57.69	2.3%	35.51 / 17.79	5.2%	74.43 / **61.34**	2.4%	**26.27** / 15.61	**21.2%**	76.56 / 77.21	2.6%	6.7%
	Soft	Task	**82.51** / 57.43	2.5%	**38.33** / 19.68	15.0%	**78.18** / 60.67	**4.4%**	23.34 / **15.89**	14.2%	77.56 / 76.55	2.8%	7.8%
		Token&Task	81.52 / 57.75	2.2%	37.54 / **20.30**	**15.8%**	77.45 / 60.42	3.7%	24.70 / 15.55	16.7%	75.61 / **76.92**	1.8%	**8.0%**
(b) Resampler design													
Compression Rate = 1			79.63 / **58.34**	1.5%	30.20 / 14.48	-12.6%	70.81 / 60.12	-1.0%	23.92 / 15.54	14.6%	**78.20** / 76.16	**3.0%**	1.1%
Compression Rate = 2, Projection			**83.74** / 57.70	**3.5%**	37.02 / 18.57	9.7%	71.89 / 60.32	-0.2%	**25.83** / 15.77	**20.5%**	74.56 / 76.70	1.0%	6.9%
Compression Rate = 4, Max Pooling			80.36 / 57.44	1.2%	27.16 / 14.37	-17.2%	68.30 / 57.56	-4.9%	18.85 / 15.60	1.1%	75.71 / 73.08	-0.7%	-4.1%
Compression Rate = 4, Avg Pooling			81.96 / 57.93	2.6%	34.21 / 14.76	-6.0%	73.39 / 59.59	0.2%	22.18 / **15.88**	11.0%	72.42 / 74.54	-1.9%	1.2%
Compression Rate = 4, Projection			81.52 / 57.75	2.2%	**37.54 / 20.30**	**15.8%**	**77.45 / 60.42**	**3.7%**	24.70 / 15.55	16.7%	75.61 / **76.92**	1.8%	**8.0%**
(c) Number of projection experts													
	3		80.45 / 56.88	0.8%	35.98 / 17.36	4.5%	66.64 / 58.10	-5.6%	24.00 / 16.00	16.3%	74.86 / 74.59	-0.3%	3.1%
	5		81.52 / 57.75	2.2%	37.54 / **20.30**	**15.8%**	**77.45** / 60.42	3.7%	24.70 / 15.55	16.7%	75.61 / 76.92	1.8%	**8.0%**
	8		82.71 / 57.86	3.0%	36.66 / 18.34	8.5%	71.15 / 58.40	-2.3%	24.47 / **15.74**	16.7%	**77.91** / 76.53	3.0%	5.8%
	10		**83.21** / 57.85	3.3%	**38.70** / 19.01	13.5%	75.06 / 61.43	2.9%	25.02 / 15.05	16.0%	77.66 / **77.99**	**3.9%**	7.9%
	16		82.92 / **58.70**	**3.9%**	35.74 / 17.66	5.1%	76.74 / **61.45**	4.1%	**27.18** / 15.48	**23.2%**	75.86 / 77.36	2.3%	7.7%
(d) Module generalization under LoRA rank setting													
rank	Connector&Router												
4	MLP		80.70 / 56.42		36.35 / 16.32		64.34 / 57.02		22.70 / 15.54		71.62 / 74.06		
	CMoE,Hard		81.94 / 57.77	2.0%	29.30 / 10.98	-26.1%	70.14 / 51.45	-0.4%	22.74 / 15.76	0.8%	81.95 / 80.46	11.5%	-2.4%
	CMoE,Soft		**82.63** / 57.66	2.3%	32.80 / 15.31	-8.0%	68.12 / **60.84**	6.3%	24.46 / 15.61	4.1%	76.11 / 73.84	3.0%	1.5%
8	MLP		79.81 / 56.48		35.18 / 16.26		74.54 / 58.42		18.55 / 15.50		76.26 / 73.64		
	CMoE,Hard		81.85 / **59.09**	3.6%	30.01 / 11.59	-21.7%	70.91 / 58.04	-2.8%	22.76 / 15.79	12.3%	81.55 / **81.18**	8.6%	0.0%
	CMoE,Soft		81.52 / 57.75	2.2%	37.54 / 20.30	15.8%	**77.45** / 60.42	3.7%	24.70 / 15.55	16.7%	75.61 / 76.92	1.8%	**8.0%**
16	MLP		79.10 / 56.45		32.73 / 14.81		72.65 / 57.89		24.42 / 16.09		69.18 / 75.43		
	CMoE,Hard		81.38 / 57.70	2.5%	30.01 / 12.89	-10.6%	71.56 / 56.42	2.4%	22.05 / 15.69	-6.1%	81.75 / 79.83	12.0%	-0.8%
	CMoE,Soft		82.54 / 58.85	**4.3%**	**38.11** / 19.13	**22.8%**	71.99 / 59.99	1.4%	26.52 / 15.58	2.7%	76.51 / 75.99	5.7%	7.4%
32	MLP		79.23 / 56.50		33.50 / 16.13		72.04 / 58.78		18.67 / 15.42		71.67 / 72.69		
	CMoE,Hard		82.25 / 58.54	3.7%	30.56 / 11.97	-17.3%	73.16 / 59.70	1.6%	23.22 / 15.58	12.7%	**82.19 / 80.67**	**12.8%**	2.7%
	CMoE,Soft		82.39 / 57.18	2.6%	35.95 / 17.03	6.4%	70.14 / 59.66	-0.6%	**26.56** / 15.45	**21.2%**	77.61 / 77.18	7.2%	7.4%
64	MLP		79.35 / 57.23		35.22 / 17.95		72.46 / 56.65		22.29 / 14.90		71.12 / 75.84		
	CMoE,Hard		81.52 / 58.55	2.5%	31.93 / 12.28	-20.5%	66.82 / 46.29	-13.0%	23.88 / 15.85	6.7%	82.00 / 79.97	10.4%	-2.8%
	CMoE,Soft		81.53 / 58.04	2.1%	35.64 / 18.81	3.0%	73.82 / 60.26	4.1%	25.89 / **16.77**	14.3%	75.21 / 77.29	3.8%	5.5%
(e) Module generalization under LoRA-MoE setting (rank = 4)													
LLM fine-tuning	Connector&Router												
LoRA	MLP		80.70 / 56.42		36.35 / 16.32		64.34 / 57.02		**22.70 / 15.54**		71.62 / 74.06		
LoRA-MoE	MLP		**85.17** / 61.29	7.1%	32.40 / 14.91	-9.7%	78.68 / 65.77	18.8%	11.26 / 14.05	-30.0%	76.66 / 78.67	6.6%	-1.4%
LoRA-MoE	CMoE,Hard		84.10 / 61.25	6.4%	31.56 / 12.64	-17.9%	78.58 / 62.51	15.9%	22.67 / 13.23	-7.5%	**80.65 / 79.89**	**10.2%**	1.4%
LoRA-MoE	CMoE,Soft		84.92 / **61.66**	**7.3%**	**39.33** / 17.10	6.5%	**79.90 / 67.69**	**21.4%**	18.73 / 13.75	-14.5%	78.90 / 77.98	7.7%	**5.7%**

hard router) and synergy hypothesis (linear, MLP, CMoE with constant router). Though the hard router has a obvious lead on the CLS task, implying that the CLS task is better suited to a separate connector to avoid conflicts with other tasks. The soft router achieves the best multi-task performance, indicating that it not only alleviates conflicts between tasks, but also promotes collaboration between tasks. We then discuss three types of router strategy. The strategy of combining token-level with task-level information is superior to using each information separately, indicating the effectiveness for considering the tug-of-war problem from both token and task level.

Resampler design   We explore whether aggregating visual features through resampler has unfavorable effects in Table 2 (b). Despite an increase in compression rate α from 1 to 4, the performance of models utilizing projection aggregation is improved. While the performance of average pooling and max pooling approaches is not satisfactory, especially the latter has severe performance degradation, which may be attributed to the excessive loss of feature information. This phenomenon shows that appropriate visual feature compression can bring efficiency to the training process without losing or even improving performance.

Number of projection experts   The number of projection experts N is one of the most significant hyperparameters, which is closely related to the number of tasks and modalities that the CMoE module can accommodate. It is a challenging study as the complexity of the scenario can end up overfitting to simpler tasks and modalities or underfitting complex ones. As shown in Table 2 (c), increasing the number of experts N, namely an augmentation in parameters, still brings performance gains on some datasets, but the average gain tends to stabilize across all tasks and datasets. Therefore, CMoE with 5 projection experts is sufficient to handle the tug-of-war problem in the existing medical multi-task learning scenarios and training configuration. A higher value of N does not bring the desired further improvement in total Δ.

4.2.2   Ablation on module generalization

We demonstrate the generalization capability of the CMoE module in any configuration, especially when the key hyperparameters and strategies for LLM fine-tuning change. We first focus on the rank of LoRA, which directly determines the LLM capacity, i.e., trainable parameters. Our observations in Table 2 (d) reveal that CMoE with soft router can steadily improve multi-task performance when LoRA rank increases from 4 to 64. In Table 2 (e), we introduce MoE to LoRA, namely LoRA-MoE, which is considered a favorable parameter-efficient tuning solution for multi-task applications [Liu *et al.*, 2023b; Chen *et al.*, 2024]. See details of LoRA-MoE at Appendix A.2. We find that separate LoRA-MoE results in significant performance improvement in 3 tasks while degradation in 2 tasks, indicating that it does not achieve the efficient solution to the tug-of-war problem. After combining CMoE with soft router, we achieve a balance of consistent performance gains, further demonstrating the necessity and effectiveness of mitigating the tug-of-war problem at the connector level in MLLMs.

4.3   Interpretation

We conduct interpretation analysis of the tug-of-war problem based on methods mentioned in Section 3.1.1. Specifically, we focus on the changes in the connector using CMoE compared to MLP and show how the tug-of-war problem is optimized: (1) From the perspective of gradient optimization, we use maximum normalization to make the tug-of-war indexes comparable under different architectures. CMoE results in a more consistent tug-of-war indexes, i.e. higher mean and smaller standard deviation, among different tasks or datasets, implying each individual gets a more balanced optimization, as shown in Figure 4 (a). (2) From the perspective of parameter statistics, we discrete the statistics scores into ten intervals and count the ratio of all parameters at connector by interval. CMoE results in an increase in the proportion of high-value intervals in Figure 4 (b). We show the routing weights of projection experts after the warm-up stage and the final model in Figure 4 (c). CMoE adaptively learns different patterns of routing weights for different tasks.

To better reflect the coexistence of conflict and synergy among tasks, as well as the critical role played by the connector, we visualize the distribution of visual features before and after passing through the connector using the t-SNE method [Van der Maaten and Hinton, 2008]. From the perspective of multi-task learning, we randomly select 200 samples from each task. It can be observed that CMoE promotes the optimization of the tug of war problem when aligning the visual space with the textual space of the LLM in Figure 5. Specifically, visual features of the same task are more

tightly distributed. For fine-grained REC and REG tasks, the distribution is highly overlapping, which facilitates synergy between tasks. For coarse-grained CLS task, the distribution is significantly different from other tasks, which is consistent with the conclusion in Section 4.2.1. We also provide visualization analysis of visual features on different medical image modalities in Appendix D.4
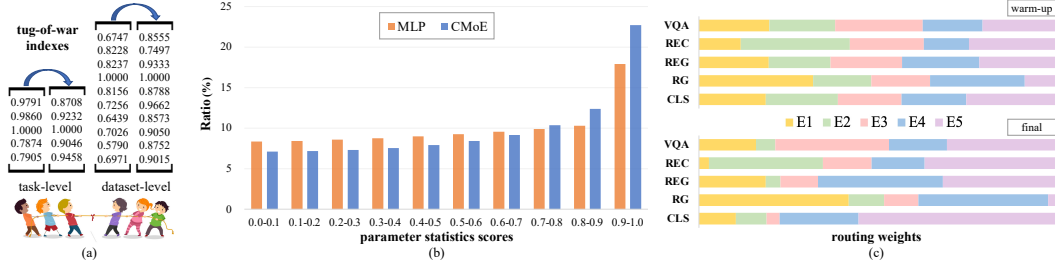

Figure 4: Interpretation analysis of the tug-of-war problem. (a) changes in tug-of-war indexes, (b) changes in the distribution of parameter statistics scores, (c) routing weights for different tasks.
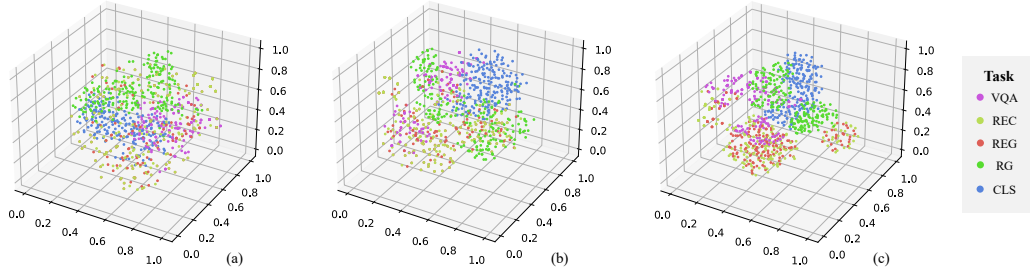

Figure 5: Visual features distribution maps-3D. (a) f_v^{ag} distribution, (b) f_v^{align} distribution obtained through MLP, (c) f_v^{align} distribution obtained through CMoE.

Table 3: Model capability comparison with open source medical MLLMs. The mean and standard deviation of performance of Uni-Med are obtained after several 300k iterations. Results with **bold**, <u>underlines</u> and gray background are the overall best, second, and zero-shot performance, respectively.

Task	Dataset	Metric	Med-Flamingo	RadFM	LLaVA-Med	XrayGPT	Uni-Med
Visual Question Answering	Slake-VQA	BLEU-1	21.51	<u>81.66</u>	76.95	-	**82.12**±0.38
		F1	23.66	<u>82.38</u>	77.30	-	**83.07**±0.34
	Path-VQA	BLEU-1	33.38	24.83	<u>46.42</u>	-	**58.07**±0.32
		F1	34.01	25.20	<u>47.08</u>	-	**58.74**±0.33
Report Generation	MIMIC-CXR	BLEU-1	23.25	6.81	19.90	<u>27.11</u>	**27.79**±2.50
		BLEU-4	1.92	1.52	0.59	<u>3.56</u>	**6.46**±0.20
		ROUGE-1	18.73	16.81	15.65	<u>24.35</u>	**28.81**±1.22
		ROUGE-2	2.28	4.48	1.13	<u>4.97</u>	**9.62**±0.99
		ROUGE-L	12.25	12.67	10.29	<u>16.29</u>	**22.58**±2.86
		METEOR	7.95	5.32	5.47	<u>9.71</u>	**10.59**±0.87
		RadGraph-F1	7.15	7.19	2.86	<u>9.00</u>	**13.98**±2.45
		RadCliQ-v0↓	4.44	4.43	4.79	<u>4.42</u>	**3.75**±0.17
		RadCliQ-v1↓	1.80	1.82	2.03	<u>1.79</u>	**1.38**±0.11
	MPx-Single	BLEU-1	8.14	-	<u>9.46</u>	8.51	**15.80**±0.24
		BLEU-4	0.45	-	<u>0.59</u>	0.23	**2.47**±0.08
		ROUGE-1	<u>11.37</u>	-	11.31	8.00	**14.32**±0.03
		ROUGE-2	0.93	-	<u>1.02</u>	0.45	**2.68**±0.01
		ROUGE-L	<u>9.65</u>	-	8.96	6.48	**12.29**±0.04
		METEOR	4.31	-	<u>5.51</u>	3.60	**5.92**±0.07
		RadGraph-F1	1.85	-	<u>2.63</u>	1.32	**4.91**±0.31
		RadCliQ-v0↓	4.00	-	<u>3.88</u>	4.17	**3.59**±0.02
		RadCliQ-v1↓	1.62	-	<u>1.55</u>	1.72	**1.37**±0.01
Image Classification	DermaMNIST	Accuracy	1.15	<u>5.14</u>	-	-	**76.96**±0.46
	OrganMNIST	Accuracy	8.90	<u>18.90</u>	-	-	**78.07**±1.63

9

4.4 Overall comparison

To demonstrate the model capabilities of Uni-Med on multi-task learning, four open source and state-of-the-art medical MLLMs including Med-Flamingo [Moor *et al.*, 2023b], RadFM [Wu *et al.*, 2023b], LLaVA-Med [Li *et al.*, 2024], and XrayGPT [Thawkar *et al.*, 2023] are used for performance comparison in Table 3. Any method of fine-tuning will inevitably lead to changes in the initial capability of the model. Therefore, we use readily available model checkpoints for testing, following the prompt template requirements of different models. Under this comparison strategy, if the training datasets of a model and Uni-Med intersect and strictly follow the official partition, it is fair and comparable to Uni-Med on these datasets. Specifically, LLaVA-MED provides dataset-specific fine-tuning checkpoints on Slake-VQA and Path-VQA separately. XrayGPT focuses on the task of report generation and utilizes MIMIC-CXR as training dataset. RadFM provides a model checkpoint for joint fine-tuning on Slake-VQA, MIMIC-CXR and MPx-Single. However, we do not list performance of RadFM on MPx-Single as we have identified the issue of data leakage, see Appendix D.2.

The results in Table 3 show that our Uni-Med achieves leading and competitive evaluation metrics across all tasks, which has the following prominent advantages: (1) Uni-Med is able to handle a greater variety of medical tasks, which is attributed to multi-task learning during training process. Due to the fact that the above MLLMs do not support input and output in coordinate form, we report the performance of Uni-Med on REC and REG tasks at Appendix D.5. Based on the different input and output forms supported by each model, we have also listed the zero-shot results in Table 3 for reference only. (2) Uni-Med achieves better results through joint training fine-tuning rather than dataset-specific fine-tuning like LLaVA-Med, which benefits from efficient optimization of the tug-of-war problem. In addition to directly compare the capability of existing models, we take LLaVA-Med as an example to compare the capability of model architectures in Appendix D.6.

5 Conclusion

In this paper, we present a novel open-source medical generalist foundation model Uni-Med, which can handle six different medical tasks including question answering, visual question answering, report generation, referring expression comprehension, referring expression generation and image classification. Benefiting from the proposed CMoE, which combines MoE with the connector, Uni-Med achieves efficient solution to the tug-of-war problem in multi-task learning. Uni-Med not only achieves competitive or superior performance compared to the open-source state-of-the-art medical MLLMs, but also provides interpretability analysis from the perspective of gradient optimization and parameter statistics on how the tug-of-war problem is optimized. We hope Uni-Med can greatly promote the development of medical generalist foundation models and inspire more research toward generalist medical artificial intelligence.

6 Limitations

While Uni-Med has demonstrated strong potential as a unified and generalist medical foundation model, it still exhibits several limitations: (1) Limitations in handling genuine 3D medical image inputs. Most commonly used medical image are in 3D. Same as most medical MLLMs, we process 3D images into 2D slices as input, resulting in significant information loss. (2) The potential of performance gains in more complex multi-modal and multi-task learning scenarios has not yet been explored. Uni-Med use 12 datasets of 6 medical tasks, with a total data volume of 140k. (3) The potential of performance gains in different LLM backbones has not yet been explored. Uni-Med utilizes LLaMA2-7B. (4) Deeper theoretical analysis of tug of war problem remains to be explored. We attempt to combine the existing methods to analyze it from the perspective gradient optimization and parameter statistics. (5) Potential negative societal impacts. We cannot prevent potential malicious or unintended uses, such as generating fake profiles or wrong medical diagnoses, and provide necessary safeguards.

Acknowledgements

References

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.

Ahmed Agiza, Marina Neseem, and Sherief Reda. Mtlora: A low-rank adaptation approach for efficient multi-task learning. *arXiv preprint arXiv:2403.20320*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. *Advances in Neural Information Processing Systems*, 33:2039–2050, 2020.

Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023.

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

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. Octavius: Mitigating task interference in mllms via moe. *arXiv preprint arXiv:2311.02684*, 2023.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11828–11837, 2023.

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

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

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

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

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

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. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

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*, pages 19730–19742. PMLR, 2023.

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, 2024.

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

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 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, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *arXiv preprint arXiv:2310.18339*, 2023.

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

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

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)*, pages 353–367. PMLR, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Shezheng Song, Xiaopeng Li, and Shasha Li. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*, 2023.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*, 2024.

Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE transactions on medical imaging*, 38(8):1885–1898, 2019.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*, 2024.

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

Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

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

Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022.

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 Component design

A.1 Type of the routing network

Constant router   The simplest routing network is to assign equal weights to the output of each expert, which can be expressed as:

$$R_{constant}(x_i) = \{1/N\}_{k=1}^{N} \tag{11}$$

Hard router   Each token is assigned to a specific expert based on its type (task / modal), with the number of experts being equal to the number of token types. It can be formulated as:

$$R_{hard}(x_i) = \{ \text{IsType}(x_i, k) \}_{k=1}^{N}$$
$$\text{IsType}(x_i, k) = \begin{cases} 1, & \text{if } x_i \text{ belongs to type } k \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Sparse router   Using a small network g, the sparse router computes a score vector for each token, with a length equal to the number of experts N. Subsequently, the Top-K function retains the top-K values in the vector, while setting all other values to zero. Finally, the *Softmax* function is applied to obtain the final routing vector. The whole process is shown as follows:

$$R_{sparse}(x_i) = Softmax(\text{Top-}K(g(x_i), K))$$
$$\text{Top-}K(v, K) = \begin{cases} v, & \text{if } v \text{ is in the top } K \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Soft Router   Similar to the sparse router, the soft router computes a score vector for each token through a small network g. Subsequently, it applies the *Sigmoid* function to the score vector and normalizes it, yielding the final routing vector. It can be formulated as:

$$R_{soft}(x_i) = \frac{Sigmoid(g(x_i))}{Sum(Sigmoid(g(x_i)))} \tag{14}$$

A.2 LoRA-MoE

LoRA-MoE freezes the original parameters of the model to preserve world knowledge and introduces LoRA experts to learn new knowledge, thereby improving performance across multiple downstream tasks with few parameters.

Specifically, given a frozen linear layer with a weight matrix $W_0 \in \mathbf{R}^{d_{in} \times d_{out}}$, LoRA-MoE creates N low-rank trainable matrix pairs A_k and B_k, where $A_k \in \mathbf{R}^{d_{in} \times r}$, $B_k \in \mathbf{R}^{r \times d_{out}}$, and the rank $r \ll min(d_{in}, d_{out})$. As in the case of LoRA, A_k is initialized with a random Gaussian distribution, and B_k is initialized to zero. During training, the parameters of W_0 are frozen, and the parameters of A_k and B_k are updated. The forward process of a LoRA-MoE layer can be represented as:

$$h = W_0 x_i + \Delta W x_i = W_0 x_i + \frac{\alpha}{r} \sum_{k=1}^{N} R(x_i) A_k B_k x_i \tag{15}$$

where x_i is the input token, R is the router in the LoRA-MoE layer, α is the learning rate scaling factor, and h is the output token. In ablation experiments, we transform each linear layer in the LLM into a LoRA-MoE layer with a sparse router. The rank $r = 4$, the learning rate scaling factor $\alpha = 8$, the number of LoRA experts $N = 5$, and select the top 2 experts.

B Dataset

B.1 Data source

MedQA   MedQA [Jin *et al.*, 2021] is a open-domain multiple-choice question answering dataset for solving medical problems. These questions are sourced from professional medical board exams, which feature diverse content and typically demand a comprehensive understanding of related medical concepts learned from medical textbooks in order to provide accurate answers. This dataset covers three languages: English, simplified Chinese, among which there are 12,723 QA pairs for English.

PubMedQA   PubMedQA [Jin *et al.*, 2019] is a biomedical question answering dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe using the corresponding abstracts. It has 1K expert-annotated, 61.2K unlabeled and 211.3K artificially generated QA instances. Each instance consists of: (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding abstract without its conclusion,(3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion.

Slake-VQA   Slake-VQA [Liu *et al.*, 2021] is a semantically annotated, knowledge-enhanced bilingual (English and Chinese) VQA dataset for radiology images. It contains 642 annotated images accompanied by 14,028 question-answer pairs, spanning 12 diseases, 39 organ systems, and 3 imaging modalities (CT, MRI, and X-ray). Questions are either open-ended (free-form) or closed-ended (balanced yes/no) related to various aspects of the image content such as plane, quality, position, organ, abnormality, size, color, shape, and knowledge graph.

Path-VQA   Path-VQA [He *et al.*, 2020] is a pathology VQA dataset comprising 4,998 pathology images and 32,799 question-answer pairs. These pathology images are sourced from medical textbooks and online digital libraries. Each image is associated with multiple QA pairs pertaining to different aspects of the pathology including color, location, appearance, shape, etc. The dataset includes 16,465 open-ended questions, which make up 50.2% of the total and are categorized into six types: what, where, when, whose, how, and how much/how many. The remaining questions are close-ended "yes/no" questions, with a balanced distribution of 8,145 "yes" answers and 8,189 "no" answers. In the official dataset split, the training set, validation set and test set contain 19,755, 6,279 and 6,761 QA pairs, respectively.

SA-Med2D-20M   SA-Med2D-20M [Ye *et al.*, 2023a] is a large-scale segmentation dataset of 2D medical images built upon numerous public and private datasets. It consists of 4.6 million 2D medical images and 19.7 million corresponding masks, covering almost the whole body and showing significant diversity. It comprises 10 modalities, with CT and MR modalities being predominant in both the number of images and masks. Specifically, there are 2338,753 images and 12547,037 masks for CT and 2217,633 images and 7147,784 masks for MR. This is primarily attributed to their widespread presence in public medical image segmentation datasets and the 3D dimension of CT and MR scans, which yields a high volume of slices when segmented across three axes.

MIMIC-CXR   MIMIC-CXR [Johnson *et al.*, 2019] is a large dataset of chest radiographs with free-text radiology reports. A total of 377,110 images are available in the dataset from 227,835 image studies collected for 65,379 patients. Each patient may have multiple studies and each study may contain one or more images associated with the same free-text report. Images in MIMIC-CXR are collected from multiple view positions: e.g., anterior-posterior (AP), posterior- anterior, and lateral (LA). Protected health information (PHI) in radiology reports and images is removed, which results in missing information in some sentences of the reports.

The MIMIC-CXR-JPG dataset is derived from MIMIC-CXR, providing JPG format files derived from the DICOM images and structured labels derived from the free-text reports. The aim of MIMIC-CXR-JPG is to provide a convenient processed version of MIMIC-CXR, as well as to provide a standard reference for data splits and image labels.

RadFM [Wu *et al.*, 2023b] processes radiology reports in MIMIC-CXR by extracting the indication, findings, and impression sections, and removing redundant white spaces. Images without reports and reports where the findings section can not be extracted are discarded from both the training and test sets. Additionally, reports with findings sections exceeding 800 characters are filtered out. To enhance the model's capability to process images from different view positions, images of different orientations associated with the same report are treated as independent samples.

MPx   MPx [Wu *et al.*, 2023b] is a report generation dataset collected from the MedPix website (https://medpix.nlm.nih.gov/) and organized by cases. Each case includes multiple radiologic scans, general clinical findings, discussions, and diagnostic results. Additionally, MPx provides scan-level annotations, such as image modality, shooting plane, and captions for each scan. The dataset is divided into MPx-Single and MPx-Multi, with annotations provided at the case level and scan level, respectively.

MedMNIST v2   MedMNIST v2 [Yang *et al.*, 2023] is a large-scale MNIST-like collection of standardized biomedical images, including 2D datasets with resolutions up to 224×224 pixels and 3D datasets with resolutions up to 64×64×64 voxels. The 2D datasets include 12 subsets: PathMNIST, ChestMNIST, DermaMNIST, OCTMNIST, PneumoniaMNIST, RetinaMNIST, BreastMNIST, BloodMNIST, TissueMNIST, OrganAMNIST, OrganCMNIST, and OrganSMNIST. The 3D datasets comprise 6 subsets: OrganMNIST3D, NoduleMNIST3D, FractureMNIST3D, AdrenalMNIST, VesselMNIST3D, and SynapseMNIST3D. Covering primary data modalities in biomedical images, it is designed to perform classification on lightweight 2D and 3D images with various data scales (from 100 to 100,000) and diverse tasks (binary/multi-class, ordinal regression and multi-label). The comprehensive dataset, comprising approximately 708K 2D images and 10K 3D images, supports a wide range of research and educational purposes in biomedical image analysis, computer vision, and machine learning.

DermaMNIST, a 2D subset of MedMNIST v2, is based on HAM10000 [Tschandl *et al.*, 2018; Codella *et al.*, 2019], a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Comprising 10,015 dermatoscopic images, the dataset is categorized into 7 distinct classes: actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.

OrganSMNIST, another 2D subset of MedMNIST v2, is based on 3D computed tomography (CT) images from Liver Tumor Segmentation Benchmark (LiTS) [Bilic *et al.*, 2023]. Organ labels are obtained by using bounding-box annotations of 11 body organs from another study [Xu *et al.*, 2019]. Hounsfield-Unit (HU) of the 3D images are transformed into grey scale with a abdominal window. Subsequently, 2D images are cropped from the center slices of the 3D bounding boxes in sagittal views. Comprising 25,211 images, the dataset is categorized into 11 distinct classes: bladder, left femur, right femur, heart, left kidney, right kidney, liver, left lung, right lung, pancreas, and spleen.

Custom dataset splitting   To prevent the model from encountering training images during testing, the official dataset split from Slake-VQA is not utilized. Instead, we randomly divide all images into training and testing sets at a ratio of 6:1, along with their respective QA pairs and bounding boxes. Consequently, the training set comprises 550 images, 6018 English QA pairs, and 1421 bounding boxes, while the testing set includes 92 images, 1014 English QA pairs, and 201 bounding boxes.

For MIMIC-CXR, JPG images provided in MIMIC-CXR-JPG and the corresponding reports from RadFM are used for the report generation task. The training set is a subset of the original training set, containing 9,997 samples, while the test set remains the same as the original test set, containing 3,858 samples.

B.2   Well-crafted datasets for REC and REG tasks

Slake-REC / Slake-REG   As a semantically-labeled knowledge-enhanced dataset for medical visual question answering, Slake-VQA provides bounding boxes for each object in the image. As shown in Figure 6 (a), the original format of each bounding box is $[X, Y, W, H]$. First, we convert it to the $[X_{min}, Y_{min}, X_{max}, Y_{max}]$ format. Assuming the relative size of each image is 100×100, we then normalize each coordinate value in the bounding box to fall within the range of 0 to 100.

As shown in Figure 6 (c), in the REC task, an image and object name are given to find the object's bounding box. In the REG task, an image and object bounding box are provided to identify the object's name. The Slake-REC and Slake-REG datasets are thus created.

SA-Med2D-REC / SA-Med2D-REG   Each image in the SA-Med2D-20M dataset has one or more masks, with each mask corresponding to an object. As shown in Figure 6 (b), we calculate the bounding box for each mask and normalize it to a range of 0 to 100, resulting in a bounding box for each object in the $[X_{min}, Y_{min}, X_{max}, Y_{max}]$ format.

The SA-Med2D-REC and SA-Med2D-REG datasets are organized as depicted in Figure 6 (c). 10,000 samples each are selected from the CT and MR subsets as the training set, and 2,000 samples each are selected as the test set.
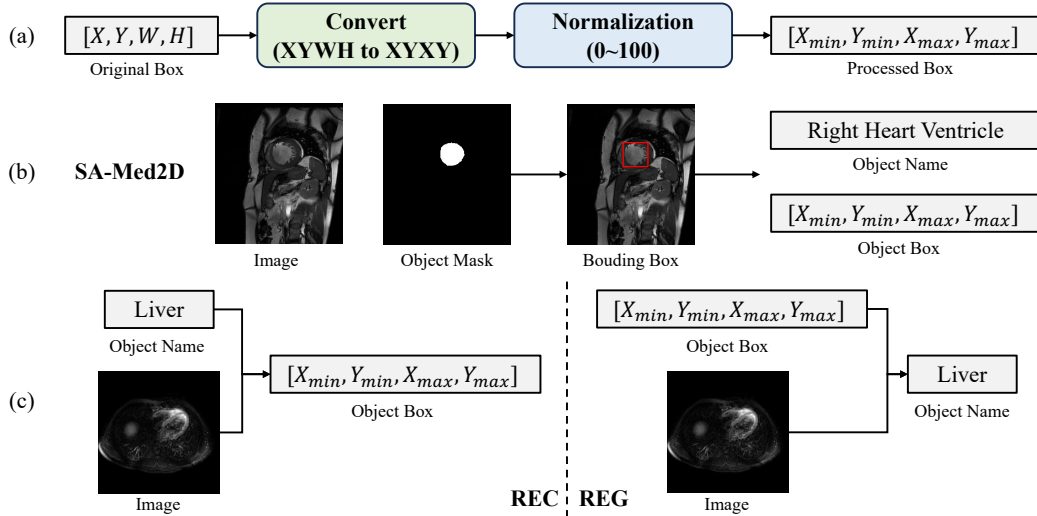
Figure 6: Data production process for REC and REG tasks. (a) the process of transforming bounding boxes in Slake-VQA, (b) the process of obtaining bounding boxes from masks in SA-Med2D, (c) the input-output organization of REC and REG tasks.

B.3 Data availability

In the Table B.3, we list the links for each dataset, the number of samples in the training and test sets, and their licenses.

Table 4: Data availability.

Dataset	Link	Train / Test Split	License
MedQA	https://github.com/jind11/MedQA	10178 / 1273	MIT License
PubMedQA	https://github.com/pubmedqa/pubmedqa	500 / 500	MIT License
Slake-VQA		6018 / 1014	Open Access
Slake-REC	https://www.med-vqa.com/slake	1421 / 201	-
Slake-REG		1421 / 201	-
Path-VQA	https://github.com/UCSD-AI4H/PathVQA	19755 / 6761	MIT License
SA-Med2D-20M		-	Apache-2.0 license
SA-Med2D-REC	https://openxlab.org.cn/datasets/GMAI/SA-Med2D-20M	20000 / 4000	-
SA-Med2D-REG		20000 / 4000	-
MIMIC-CXR	https://physionet.org/content/mimic-cxr-jpg/2.1.0 https://huggingface.co/datasets/chaoyi-wu/RadFM_data_csv	9997 / 3858	PhysioNet Credentialed Health Data License 1.5.0 Apache-2.0 license
MPx-Single	https://huggingface.co/datasets/chaoyi-wu/MedPix-Images https://huggingface.co/datasets/chaoyi-wu/RadFM_data_csv	31416 / 6664	Open Access Apache-2.0 license
DermaMNIST	https://medmnist.com	7007 / 2005	Apache-2.0 License
OrganSMNIST	https://medmnist.com	13932 / 8827	Apache-2.0 License

C Multi-task instruction template

We have designed different instruction templates for different datasets. During the training process, when a sample from a dataset is selected, an instruction template is also sampled from the corresponding dataset's template pool and used to format the sample. Examples of instruction templates for each dataset are shown below.

MedQA

Example 1: [qa] A researcher evaluates healthy breast tissue from 100 women, 50 women that were pregnant at the time of the study and 50 age-matched non-pregnant women. The breast tissue in pregnant women contained an increased number of acinar glands with epithelial proliferation compared to the non-pregnant women. Which process caused this change?

Example 2: [qa] If you are a doctor, please answer the following question briefly: a researcher evaluates healthy breast tissue from 100 women, 50 women that were pregnant at the time of the study and 50 age-matched non-pregnant women. The breast tissue in pregnant women contained an increased number of acinar glands with epithelial proliferation compared to the non-pregnant women. Which process caused this change?

PubMedQA

Example 1: [qa] Does the severity of obstructive sleep apnea predict patients requiring high continuous positive airway pressure?

Example 2: [qa] If you are a doctor, please answer the following question using "yes", "no" or "maybe": does the severity of obstructive sleep apnea predict patients requiring high continuous positive airway pressure?

Slake-VQA / Path-VQA

Example 1: <Img> <ImageFeature> </Img> [vqa] What modality is used to take this image?

Example 2: <Img> <ImageFeature> </Img> [vqa] Based on the image, respond to this question with a short answer: what modality is used to take this image?

Slake-REC / SA-Med2D-REC

Example 1: <Img> <ImageFeature> </Img> [refer] Liver.

Example 2: <Img> <ImageFeature> </Img> [refer] Give me the location of liver.

Example 3: <Img> <ImageFeature> </Img> [refer] Where is liver?

Example 4: <Img> <ImageFeature> </Img> [refer] From this image, tell me the location of liver.

Example 5: <Img> <ImageFeature> </Img> [refer] The location of liver is

----

Example 6: <Img> <ImageFeature> </Img> [refer] Could you tell me the location for liver?

Example 7: <Img> <ImageFeature> </Img> [refer] Where can I locate the liver?

Slake-REG / SA-Med2D-REG

Example 1: <Img> <ImageFeature> </Img> [identify] <16><36><42><61>

Example 2: <Img> <ImageFeature> </Img> [identify] What object is in this location <16><36><42><61>?

Example 3: <Img> <ImageFeature> </Img> [identify] Identify the object present at this location <16><36><42><61>.

Example 4: <Img> <ImageFeature> </Img> [identify] What is it in <16><36><42><61>?

Example 5: <Img> <ImageFeature> </Img> [identify] Describe this object in <16><36><42><61>.

Example 6: <Img> <ImageFeature> </Img> [identify] This <16><36><42><61> is

---

Example 7: <Img> <ImageFeature> </Img> [identify] The object in <16><36><42><61> is

MIMIC-CXR

Example 1: <Img> <ImageFeature> </Img> [caption] Describe the given chest x-ray image in detail.

Example 2: <Img> <ImageFeature> </Img> [caption] Take a look at this chest x-ray and describe the findings and impression.

Example 3: <Img> <ImageFeature> </Img> [caption] Could you provide a detailed description of the given x-ray image?

Example 4: <Img> <ImageFeature> </Img> [caption] Describe the given chest x-ray image as detailed as possible.

Example 5: <Img> <ImageFeature> </Img> [caption] What are the key findings in this chest x-ray image?

MPx-Single

Example 1: <Img> <ImageFeature> </Img> [caption] Describe this input image.

Example 2: <Img> <ImageFeature> </Img> [caption] Help captioning the image.

Example 3: <Img> <ImageFeature> </Img> [caption] What can be inflected from the scan?

Example 4: <Img> <ImageFeature> </Img> [caption] Can you give a caption for this image?

Example 5: <Img> <ImageFeature> </Img> [caption] Can you provide a brief summary of the radiology image?

Example 6: <Img> <ImageFeature> </Img> [caption] Please write a report about the image?

Example 7: <Img> <ImageFeature> </Img> [caption] Can you provide an analysis of this image?

Example 8: <Img> <ImageFeature> </Img> [caption] Can you explain what is shown in this image?

Example 9: <Img> <ImageFeature> </Img> [caption] What can be indicated from the radiologic scans?

Example 10: <Img> <ImageFeature> </Img> [caption] What can you infer from this photograph?

> **DermaMNIST**
>
> **Example**: <Img> <ImageFeature> </Img> [cls] Which category does this multi-source dermatoscopic image of common pigmented skin lesions belong to: actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, or vascular lesions?

> **OrganSMNIST**
>
> **Example**: <Img> <ImageFeature> </Img> [cls] Which category does this CT image belong to: bladder, left femur, right femur, heart, left kidney, right kidney, liver, left lung, right lung, pancreas, or spleen?

D Experiments

D.1 Evaluation metrics

F1 Score Assuming m is the number of common words in the candidate C and the reference R with the number of words of c and r, the precision and recall for a candidate sentence can be calculated as:

$$precision = \frac{m}{c} \tag{16}$$

$$recall = \frac{m}{r} \tag{17}$$

Considering class imbalance, F1 score is used to evaluate the performance of the model on both the VQA and REG tasks, which means the harmonic mean of precision and recall. A higher average F1 score for the dataset indicates a higher performance of the model.

$$\text{F1} = \frac{2 \times precision \times recall}{precision + recall} \tag{18}$$

BLEU-N We use BLEU-1 to assess the model's performance on both the VQA and REG tasks, while employing both BLEU-1 and BLEU-4 to evaluate its performance in the report generation task. Given the candidate C and reference R, BLEU-N is defined as:

$$\text{BLEU-N} = \frac{\sum_{\text{gram}_N \in C} Count_{clip}(\text{gram}_N)}{\sum_{\text{gram}_N \in C} Count(\text{gram}_N)} \tag{19}$$

When N=1, the above formula calculates BLEU-1; when N=4, it calculates BLEU-4.

ROUGE-N We use ROUGE-1 and ROUGE-2 to evaluate the performance of the model on the RG task. Given the candidate C and reference R, ROUGE-N is defined as:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_N \in R} Count_{match}(\text{gram}_N)}{\sum_{\text{gram}_N \in R} Count(\text{gram}_N)} \tag{20}$$

When N=1, the above formula calculates ROUGE-1; when N=2, it calculates ROUGE-2.

ROUGE-L ROUGE-L is also used to evlaute the quality of the generated text on the task of report generation, which stands for recall-oriented understudy for gisting evaluation with the longest common subsequence. Given the candidate C and reference R, let $LCS(C, R)$ be the length of the longest common subsequence, which is determined by using dynamic programming, it can be an defined as:

$$\text{ROUGE-L} = \frac{(1 + \beta^2) R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{21}$$

where $R_{LCS} = \frac{LCS(C,R)}{L_C}$, $P_{LCS} = \frac{LCS(C,R)}{L_R}$, $\beta = \frac{P_{LCS}}{R_{LCS}}$. L_C and L_R represent the length of the candidate and reference. A higher ROUGE-L score means that the generated text shares more of the same sequences of words as the reference text, which typically indicates better quality in terms of capturing the salient points of the reference.

METEOR METEOR is also used to evlaute the quality of the generated text on the task of report generation, which stands for metric for evaluation of translation with explicit ordering. METEOR for a sentence is computed as:

$$\text{METEOR} = (1 - p) \times \frac{precision \times recall}{\alpha \times precision + (1 - \alpha) \times recall} \tag{22}$$

where $p = \gamma(\frac{ch}{m})^\theta$ is the penalty factor. ch is the number of chunks, which means a contiguous ordered block. α, θ and γ are hyperparameters determined according to different datasets.

RadGraph F1 To assess the semantic accuracy in the task of report generation, RadGraph F1 computes the overlap in clinical entities and relations between a machine-generated report and a radiologist-generated report. Specifically, following the criteria in RadGraph [Jain *et al.*, 2021], two entities are matched if their tokens (words in the original report) and labels (entity type) match. Two relations are matched if their start and end entities match and the relation type matches. RadGraph F1 metric computes the overlap in entities and relations separately and reports their average.

RadCliQ RadCliQ (radiology report clinical quality) is also used to assess the semantic accuracy in the task of report generation. Two versions of the RadCliQ metric: RadCliQ-v0 and RadCliQ-v1 both use a machine learning model to take in values from other metrics, such as BERTScore and CheXbert vector similarity, and then produce a composite score based on these input values, which predict the total number of errors in a report.

IoU We use IoU (Intersection over Union) to evaluate the performance of the model on the REC task. It can be formulated as:

$$\text{IoU} = \frac{P \cap G}{P \cup G} \tag{23}$$

where P is the prediction area of the model, G is the area of the ground truth.

R@0.5 We alse use R@0.5 to evaluate the performance of the model on the referring expression comprehension task. R stands for recall, and 0.5 denotes the IoU threshold. When the IoU between the prediction and the ground truth is greater than or equal to 0.5, it is considered a true positive (TP). When the IoU is less than 0.5, it is considered a false negative (FN). Therefore, for a sample with only one bounding box, R@0.5 can be formalized as:

$$\text{R@0.5} = \frac{TP}{TP + FN} = \begin{cases} 1, & \text{IoU} \geq 0.5 \\ 0, & \text{IoU} < 0.5 \end{cases} \tag{24}$$

D.2 Data leakage issue of RadFM on MPx-Single

When we directly use the model checkpoint provided by RadFM open-source repository for model inference, we find that the model outputs for many samples were completely consistent with ground truth. This issue only occurs on the MPx-Single dataset, while we strictly follow the official test set split. This appears to be unreasonable, raising suspicions of potential data leakage. Here are some examples:

> **Example 1**
>
>
>
> **Instruction**: Describe this input image.
>
> **Ground truth**: Neuroradiology Quiz
>
> **RadFM**: Neuroradiology Quiz
>
> **Uni-Med**: axial noncontrast ct image shows a large, well-circumscribed, non-aggressive appearing mass in the right frontal lobe.
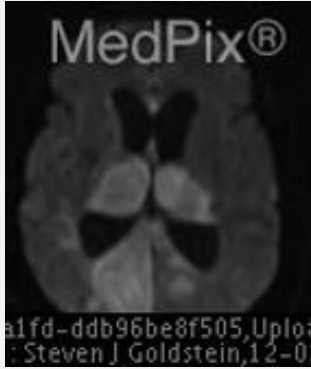
Example 2

Instruction: Describe this input image.

Ground truth: Focal necrosis of splenium of corpus callosum secondary to subacute diffuse axonal injury.

RadFM: Focal necrosis of splenium of corpus callosum secondary to subacute diffuse axonal injury.

Uni-Med: there is a large, well-circumscribed, extra-axial mass in the right frontal region. the mass is predominantly cystic with a few foci of solid tissue. there is no evidence of hemorrhage or enhancement on the mr sequences.

Example 3

Instruction: Describe this input image.

Ground truth: MRI day 2- hydrocephalus Acute bilateral cerebellar infarcts Acute bilateral thalamic infarcts Acute right occipital lobe infarct MRA no flow in distal basilar artery or posterior cerebal arteries.

RadFM: MRI day 2- hydrocephalus Acute bilateral cerebellar infarcts Acute bilateral thalamic infarcts Acute right occipital lobe infarct MRA no flow in distal basilar artery or posterior cerebal arteries.

Uni-Med: acute right mca infarct. acute infarction of the right cerebellar hemisphere. acute infarction of the right brainstem. acute cerebral edema.

D.3 Ablation on special token and identifier

We have designed vision-level special task tokens and text-level special task identifiers for visual features and text prompt, respectively. Through ablation experiment, we verify whether they have a positive effect on model performance. As shown in Table 5, we observe that text-level special task identifiers bring limited improvement. In contrast, vision-level special task tokens significantly improve the model's overall performance on all datasets, further illustrating the effectiveness of mitigating the tug-of-war problem at the connector.

Table 5: Ablation Experiments on special token and identifier.

Connector	Special Token / Identifier Text-level	Vision-level	VQA BLEU-1	Avg.	REC IoU	Δ (↑)	REG BLEU-1	Δ (↑)	RG BLEU-1	Δ (↑)	CLS Accuracy	Δ (↑)	Total Δ (↑)
MLP	-	-	79.81 56.48		35.18 16.26		74.54 58.42		18.55 15.50		76.26 73.64		
CMoE	-	-	81.59 57.35	1.9%	36.76 18.74	9.9%	76.07 58.81	1.4%	24.71 15.42	16.4%	74.46 76.07	0.5%	6.0%
	-	✔	81.33 57.29	1.7%	**37.85** 20.14	15.7%	77.23 **62.72**	5.5%	23.29 **15.74**	13.6%	**76.76** 76.55	2.3%	7.8%
	✔	-	**81.79** 57.69	**2.3%**	35.51 17.79	5.2%	74.43 61.34	2.4%	**26.27** 15.61	**21.2%**	76.56 **77.21**	**2.6%**	6.7%
	✔	✔	81.52 **57.75**	2.2%	37.54 **20.30**	**15.8%**	**77.45** 60.42	3.7%	24.70 15.55	16.7%	75.61 76.92	1.8%	**8.0%**

D.4 Visualization analysis of visual features on image modalities

We use t-SNE to visualize the distribution of visual features by modalities in Figure 7. We first observe the visual feature distribution of different modalities under the same task in Figure 7 (a-c). The feature of CT and MRI modalities in the REG task already have good discriminability after
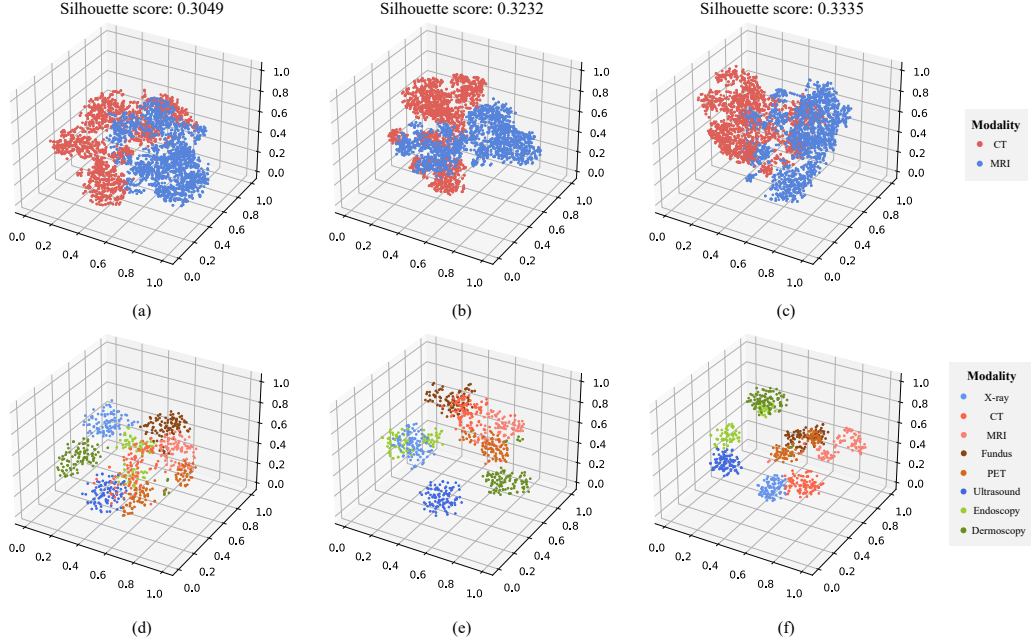
Figure 7: Visual features distribution on image modalities. (a)-(c) The feature distribution of CT and MRI modalities in the REG task. (a) passing through the frozen visual encoder.(b) passing through the MLP connector. (c) passing through the CMoE. (d)-(f) The feature distribution of 8 modalities after passing through the frozen visual encoder. (d) EVA-CLIP ViT-G/14. (e) CLIP ViT-L/14. (f) BiomedCLIP ViT-B/16.

passing through the frozen visual encoder. After passing through the connector, the improvement in Silhouette score (from 0.3049 to 0.3335) is relatively limited. In addition, we select 100 samples from each of the 8 modalities and observe their visual feature distributions after passing through different visual encoders in Figure 7 (d-f). It can still be observed that visual features of different modalities already have specific patterns in the feature space, whether using EVA-CLIP, CLIP or BiomedCLIP.

The above findings also provide an explanation for why we attempt to explicitly introduce task information instead of modality information in CMoE. When aligning visual and language embedding spaces through the connector in Uni-Med's multi-modal and multi-task scenario, task information is more difficult to distinguish than modality information.

D.5 Performance of Uni-Med on REC and REG tasks

We report the metrics of Uni-Med on the tasks of referring expression comprehension and referring exression generation in Table 6. The mean and standard deviation of performance of Uni-Med are obtained after several 300k iterations.

Table 6: Performance of Uni-Med on REC and REG tasks.

Task	Dataset	Metric	Uni-Med
Referring Expression Comprehension	Slake-REC	IoU	37.71±0.52
		R@0.5	39.30±0.76
	SA-Med2D-REC	IoU	21.60±2.19
		R@0.5	14.42±3.20
Referring Expression Generation	Slake-REG	BLEU-1	75.78±2.05
		F1	77.35±1.97
		Accuracy	68.16±1.32
	SA-Med2D-REG	BLEU-1	61.47±1.76
		F1	62.17±1.90
		Accuracy	57.69±1.07

24

D.6 Comparison of architecture capability between Uni-Med and LLaVA-Med

In addition to directly compare the capability of existing models, we take LLaVA-Med as an example to compare the capability of model architectures.

Specifically, we use the checkpoints of the second stage (medical instruction tuning) to perform two strategies of LLM full parameter fine-tuning: (1) Dataset-specific fine-tuning; (2) Joint training fine-tuning. The data split and the prompt format are completely consistent with Uni-Med and LLaVA-Med, respectively. Both strategies last for 3 epochs (the same as Uni-Med). The results are shown in Table 7.

Following the model architecture of LLaVA-Med, there is a serious tug-of-war problem when we implement joint fine-tuning strategy on multiple tasks and datasets. While the strategy of dataset-specific fine-tuning has significantly improved the evaluation metrics of each dataset.

It is worth noting that Uni-Med has achieved competitive and leading results through joint training, without dataset-specific fine-tuning. It can be concluded that the model architecture of Uni-Med, especially the design of CMoE, has achieved a superior solution to the tug-of-war problem, which reduces interference and promotes more efficient knowledge sharing.

Table 7: Comparison of architecture capability between Uni-Med and LLaVA-Med. We utilize dataset-specific fine-tuning and joint training fine-tuning on LLaVA-Med, respectively.

Task	Dataset	Metric	LLaVA-Med		Uni-Med
			Joint Training	Dataset-specific	Joint Training
Visual Question Answering	Slake-VQA	BLEU-1	33.69	72.00	**82.12**
		F1	35.83	73.07	**83.07**
	Path-VQA	BLEU-1	37.79	56.86	**58.07**
		F1	38.55	57.51	**58.74**
Report Generation	MIMIC-CXR	BLEU-1	20.43	21.03	**27.79**
		BLEU-4	4.86	4.96	**6.46**
		ROUGE-1	26.11	28.28	**28.81**
		ROUGE-2	7.66	9.01	**9.62**
		ROUGE-L	19	20.61	**22.58**
		METEOR	8.73	8.89	**10.59**
	MPx-Single	BLEU-1	15.11	14.63	**15.80**
		BLEU-4	2.4	1.75	**2.47**
		ROUGE-1	13.22	13.03	**14.32**
		ROUGE-2	2.39	2.19	**2.68**
		ROUGE-L	10.99	10.85	**12.29**
		METEOR	5.83	5.79	**5.92**
Image Classification	DermaMNIST	Accuracy	25.84	**79.95**	76.96
	OrganSMNIST	Accuracy	66.80	77.84	**78.07**
Referring Expression Comprehension	Slake-REC	IoU	4.07	22.41	**37.71**
		R@0.5	1.99	18.41	**39.30**
	SA-Med2D-REC	IoU	8.64	17.67	**21.60**
		R@0.5	4.75	9.98	**14.42**
Referring Expression Generation	Slake-REG	BLEU-1	27.21	50.79	**75.78**
		F1	30.97	53.15	**77.35**
		Accuracy	20.40	44.78	**68.16**
	SA-Med2D-REG	BLEU-1	45.83	55.15	**61.47**
		F1	47.11	55.98	**62.17**
		Accuracy	40.80	50.92	**57.69**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our contributions and scope in abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Although the paper does not include theoretical results, but uses and combines existing theoretical methods for interpretability analysis about the tug-of-war problem. All assumptions are clearly stated or referenced in the statement.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and resources are available at `https://github.com/tsinghua-msiip/Uni-Med`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided all the training and test details including data splits, hyperparameters, how they were chosen, type of optimizer, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In overall comparison, we report the mean and standard deviation of performance of our model on all tasks, which is calculated through three different random seed configurations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have provided the information about computer resources such as GPU and time of execution in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We have discussed the potential positive societal impacts in section of Introduction and Conclusion, and negative societal impacts in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: For our open-source model, providing effective safeguards is a challenge.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All creators or original owners of assets used in the paper are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have produced new datasets for several tasks and described the processing process in Appendix B.2.

    Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.