

MedRev: Reversing Negative Transfer in Medical Multi-modal Foundation Model Fine-tuning for Improved Medical Image Analysis

Yuncheng Jiang, Chun-Mei Feng, Jiankun Hu, Lusheng Wang, Le Zhang^(✉)

Abstract Medical multi-modal foundation models (MFMs) have shown strong generalization potential for smart clinical applications. However, conflicting cross-modal priors can cause them to underperform single-modal foundation models on specific downstream tasks. We define this phenomenon as “negative multi-modal transfer”, where the transfer gain of a multi-modal backbone becomes inconsistent or negative for a target modality. Through controlled experiments, we observe that (1) the extent of pretraining and (2) which layers are selected for fine-tuning substantially affect whether knowledge transfers effectively across modalities in downstream tasks. To address this, we propose MedRev, an efficient fine-tuning paradigm developed to reverse the negative transfer effect in medical foundation models. MedRev introduces a mixture-of-pre-trained experts strategy that adaptively leverages knowledge from different pre-trained checkpoints based on downstream data. Moreover, a dynamic layer freezing strategy is employed to decrease unnecessary updates and fine-tune those most beneficial for the target modality. Extensive experiments on downstream datasets across ten modalities demonstrate that MedRev consistently mitigates the negative transfer and increases fine-tuning performance.

Keywords Deep learning, Multimodal pretraining, Negative transfer, Mixture of Experts.

1 Introduction

Generalized multi-modal foundation models (MFMs) have become an increasingly important paradigm in medical image analysis, where acquiring large amounts of high-quality annotations is often infeasible [1, 2]. MFMs are pre-trained on

large, heterogeneous datasets using self-supervised [3, 4] or pretext objectives [5, 6] to learn a generic encoder that integrates signals from multiple modalities, e.g. images, text, and other clinical measurements. The pre-trained representations are then adapted to downstream clinical tasks via fine-tuning [7, 8], adapter modules [9, 10], or lightweight classifiers [5, 11]. By leveraging abundant unlabeled or weakly labeled data during pre-training, MFMs substantially reduce the annotation burden and enable more robust cross-modal reasoning.

However, the success of transfer from a pre-trained foundation model to downstream tasks is not always guaranteed. When the source and target domains exhibit sufficient gaps, transferring from such weakly correlated knowledge can actually degrade performance rather than increase target performance, a phenomenon known as negative transfer [12]. Multi-modal pre-training may further intensify this problem, as heterogeneous signals from different modalities may induce representation conflicts, spurious correlations, or misaligned feature spaces during pre-training. In this manuscript, we use modality conflicts as an umbrella term for these cross-modal discrepancies, which are reflected by lower or negative transfer gain after fine-tuning. Consequently, despite offering richer multi-modal priors, MFMs can underperform carefully designed single-modal foundation models when cross-modal interference outweighs the benefit of additional modalities.

The concept of negative transfer has been widely recognized in the field of transfer learning [13]. Recent works have observed this phenomenon in both supervised and self-supervised learning contexts [14, 15]. For instance, Wang et al. have conducted empirical studies to demonstrate the existence of negative transfer in supervised learning [16]. Ye et al. showed the performance gaps between the single and multi-modal self-supervised learning [17]. Despite these empirical observations and attempts, few studies have analyzed multi-modal negative transfer specifically within the medical field. The following issues remain unresolved: 1) What: Although the concept is relatively intuitive, there is still a lack of clarity on precisely defining multi-modal negative transfer for medical multi-modal foundation models. For example,

• Yuncheng Jiang and Jiankun Hu are with the Department of General Surgery, Gastric Cancer Center & Laboratory of Gastric Cancer, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, 610041, China.

• Chun-Mei Feng is with the School of Computer Science, University College Dublin, D04 V1W8, Ireland.

• Lusheng Wang is with the Department of Computer Science, City University of Hong Kong, 999077, Hong Kong, China.

• Le Zhang is with the College of Computer Science, Sichuan University, Chengdu, 610065, China. E-mail: zhangle06@scu.edu.cn.

* To whom correspondence should be addressed.

Manuscript received: 20xx-xx-xx; revised: 20xx-11-xx; accepted: 20xx-00-xx



what metric should we use to measure it, and what is the impact of multi-modal negative transfer on each modality? 2) Why: Why does the multi-modal negative transfer occur? Are there any factors that correspond to this issue? 3) How: Most importantly, how can we mitigate multi-modal negative transfer and boost the performance on various medical imaging modalities through an efficient fine-tuning process?

The main contributions in this paper originate from the step-by-step analysis of the above issues:

What: Multi-modal knowledge in pre-training data can yield both positive and negative effects on downstream medical imaging tasks, but the negative influence tends to dominate. In Sections 3.1 and 3.2, we derive a formal definition of the multi-modal transfer gain to quantify negative transfer for medical foundation models. We further conduct experiments to assess the impact of multi-modal negative transfer on 10 clinically relevant downstream tasks spanning **CT, MRI, ultrasound, and other modalities**. Seven out of the ten datasets exhibit measurable negative transfer (Fig. 1(b) and (c)).

Why: Intermediate pre-training checkpoints and the positioning of fine-tuning learnable layers significantly impact the overall performance and transfer gain. In Sections 3.3 and 3.4, we show that certain intermediate checkpoints of the pre-training process may positively affect the performance on downstream medical tasks (Fig. 2). Besides, we observe that fine-tuning layers contribute differently to transfer gain, with gradient changes of each layer corresponding to this pattern (Fig. 3).

How: We developed a straightforward fine-tuning paradigm to mitigate the impact of negative transfer for medical multi-modal foundation models. In Section 4, based on the above analysis, we design the mixture of pre-trained experts (MoPE) to fuse knowledge from multiple pre-training epochs. Besides, we propose dynamic layer freezing (DLF) to select learnable layers according to the gradient norm during fine-tuning (Fig. 4). In Section 5, we validate our method across ten medical downstream datasets and present its superior performance on clinical metrics and reduced negative transfer.

2 Related Work

2.1 Self-supervised Learning in Medical Foundation Model

Foundation models for medical applications have evolved along two complementary trajectories: modality-specific and modality-agnostic approaches. Modality-specific models are pretrained and optimized for a single data type (e.g., CT

[18, 19], pathology slides [20], endoscopy [21–23], or ultrasound images [24, 25]) to solve specific clinical problems. In contrast, modality-agnostic or multi-modal foundation models (MFMs) aim to embed heterogeneous inputs into a shared representation space, and then a single pretrained encoder can be adapted to a wide range of clinical tasks. Representative MFMs include MedSAM [26], BiomedGPT [27] and LVM-FM [28]. Central to both lines of work is the choice of pretraining paradigm: self-supervised learning (SSL) and related pretext objectives provide scalable means to exploit large amounts of unlabeled or weakly labeled clinical data. SSL methods in medical imaging can be broadly grouped into three categories: (1) Single-modal SSL, which learns strong within-modality features via contrastive learning [29] or masked reconstruction [5] but does not explicitly leverage cross-modal correlations. (2) Multi-modal SSL, which leverage aligned data (e.g. image-text pairs) to align heterogeneous modalities through joint embedding spaces. Wang et al. [30] pioneered CLIP to align visual and textual embeddings in medical imaging, and Yan et al. first introduce domain knowledge into BERT [31] pre-training. Yet, such paired datasets are scarce in clinical practice. (3) Unpaired multi-modal SSL, which looks for the modality-agnostic representations from non-aligned datasets. Chen et al. [32] developed universal image-text representation learning for mining efficient context relations in unpaired hidden space. **Ye et al. [17] to accommodate an increasing number of new modalities.** Although MFMs offer rich priors to boost few-shot generalization, it brings unique challenges like representation conflicts and negative transfer.

2.2 Adaptation for Foundation Models

As foundation models grow in scale, full fine-tuning becomes computationally prohibitive and prone to overfitting, especially on small medical datasets. Consequently, parameter-efficient fine-tuning (PEFT) and model adaptation strategies have gained prominence. Adapter-based [9] approaches insert lightweight bottleneck modules between existing layers, allowing the original parameters to remain frozen while learning only a small fraction of task-specific weights. Alternatively, Low-Rank Adaptation (LoRA) [33] injects trainable rank decomposition matrices into the architecture, hypothesizing that weight updates have a low intrinsic rank, thereby reducing the hardware barrier without introducing inference latency. Further advancements include Scaling and Shifting Features (SSF) [34], which modulate deep features via linear transformations that can be re-parameterized during inference for zero-overhead adaptation, and initialization strategies like

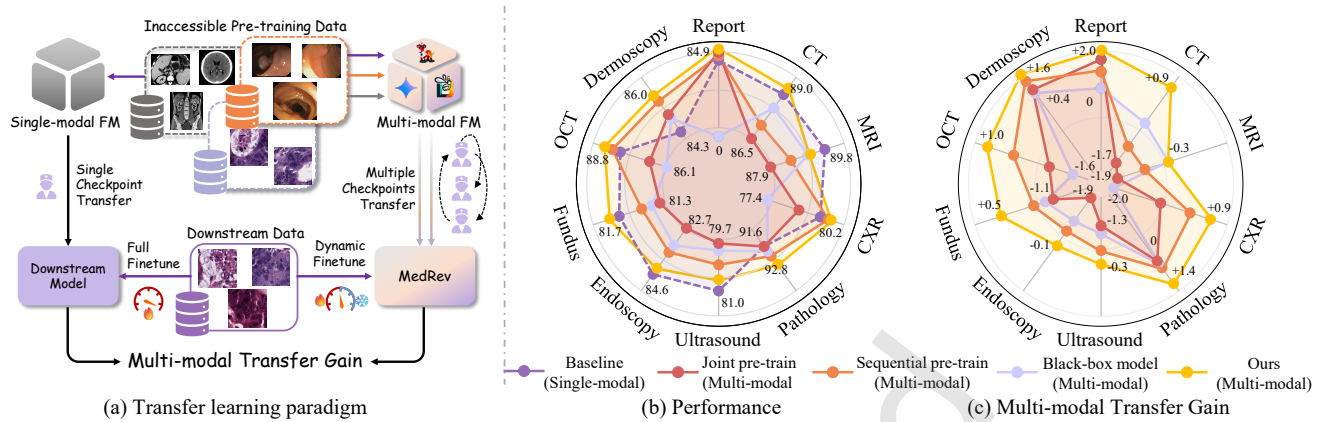


Fig. 1 (a) The transfer learning paradigm for medical multi-modal foundation models. (b) Downstream performance on 10 modality-specific datasets, illustrating that negative transfer is a noticeable problem in various modalities. (c) Multi-modal transfer gain after fine-tuning.

Target Parameter Pre-training (TPP) [35], which argues that newly added parameters should be pre-trained on target data rather than randomly initialized to fully exploit the backbone's potential. Beyond typical PEFT, other works have explored merging models from different initializations to boost transfer performance [36]. However, despite these advancements in parameter efficiency, existing PEFT methods often assume a benign transfer of knowledge; yet, in multimodal pre-training scenarios, the distinct optimization dynamics of different modalities can lead to interference, where the adaptation process inadvertently degrades representations or fails to align conflicting modal distributions. Our method specifically targets this gap by addressing the negative transfer arising during the fine-tuning of multimodal pre-trained models, ensuring that efficiency does not come at the cost of modal alignment or robustness.

2.3 Negative Transfer

Negative transfer (NT) occurs when leveraging source data or pre-trained models degrades rather than increases performance on a target task [14]. Empirical and theoretical studies have shown that negative transfer arises from several sources, including low domain similarity, inclusion of irrelevant or noisy source examples, and the injection of spurious or modality-specific features that conflict with target objectives [16]. To mitigate these risks, prior works [12, 16, 37] have pursued two complementary strategies. First, data selection and instance reweighting methods aim to suppress unhelpful source samples or emphasize target-relevant examples by boosting-based learning [38]. Second, representation-level approaches try to align source and target feature spaces or filter untransferable components during pretraining and adaptation, which include adversarial filtering and transferability-aware

feature selection [37]. Third, optimization-level remedies address harmful gradient interactions during multi-objective or multi-modal training, for example by performing gradient surgery or task-aware gradient projection to avoid destructive interference [32]. Beyond these general strategies, the challenge of NT is amplified in multi-modal models, where fine-tuning can exacerbate spurious cross-modal correlations learned during pre-training. This necessitates the design of targeted fine-tuning objectives and regularization techniques to suppress such detrimental associations [39]. The problem is particularly acute in the context of medical imaging, a domain frequently characterized by pronounced class imbalance, heterogeneous data acquisition protocols and limited availability of target datasets. Consequently, recent methods tailored for medical applications have been proposed, such as transfer-risk maps [40] proposed by Duan et al., which aim to localize and mitigate NT in dense prediction tasks like segmentation. Niu et al. propose a novel distant domain transfer learning that leverages easily accessible and unlabeled data from distant domains for mitigating NT in COVID-19 diagnosis [41]. Different from these lines of work, our study systematically investigates negative transfer (NT) in multi-modal medical foundation models by addressing three core questions: what forms of negative transfer manifest, why they emerge across different modalities, and how they can be effectively reversed.

3 Understanding the Negative Transfer in Medical Multi-Modal Foundation Models

3.1 Problem Definition

Consider a large-scale medical multi-modal dataset $\mathcal{D}_p$: $\{\mathcal{D}_p^1, \mathcal{D}_p^2, \dots, \mathcal{D}_p^M\}$, containing various datasets with M different modalities. The pre-training stage consists of a self-supervised pretext task applied to a foundation model θ_p ,

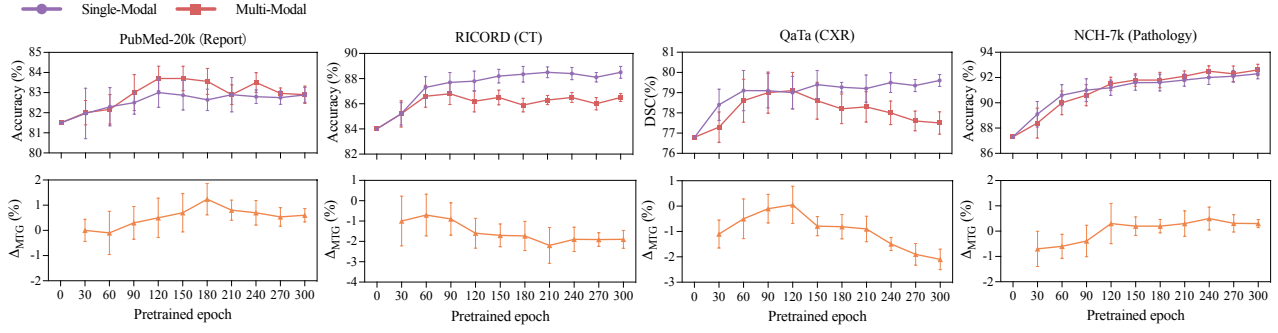


Fig. 2 Transfer learning performance (top) and multi-modal transfer gain (bottom).

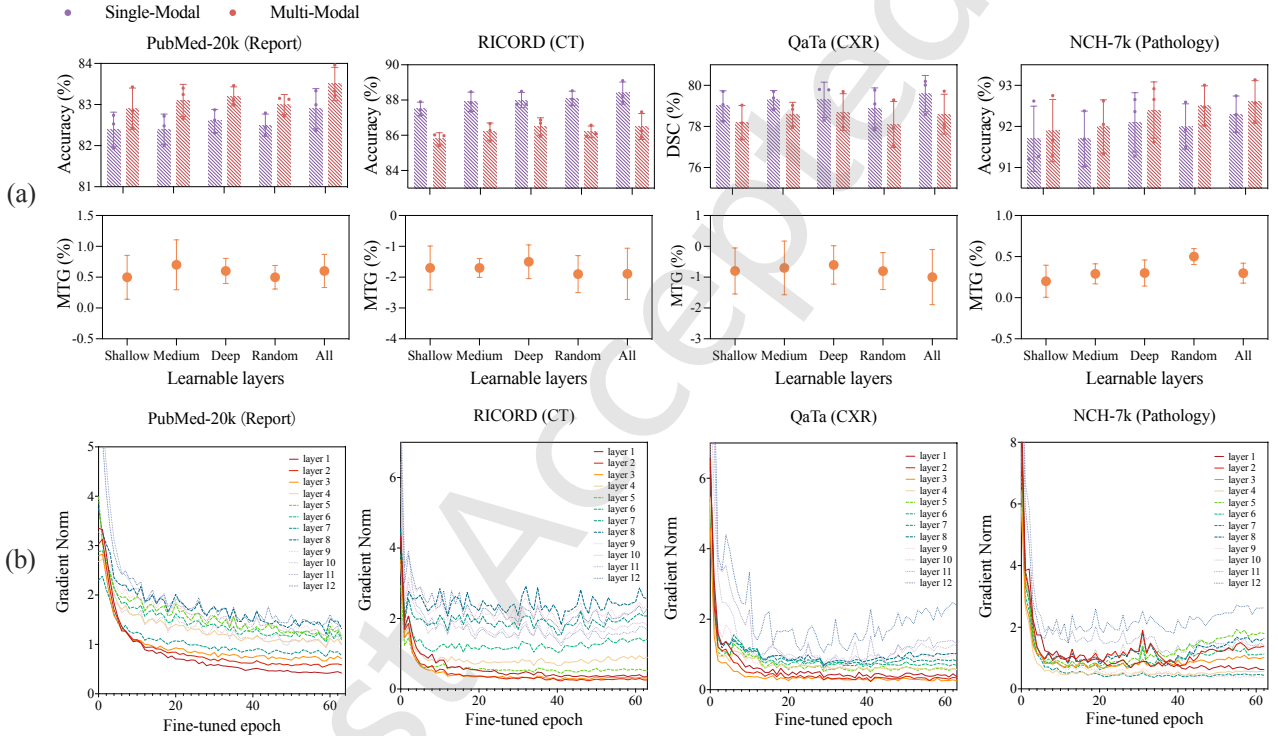


Fig. 3 (a) Transfer learning performance (top) and multi-modal transfer gain (middle). (b) Gradient norm of each layer. We can observe that the medium and deep layers typically contribute more to transfer learning performance.

allowing it to learn generalizable features from unlabeled multi-modal data. In addition, each modality has a corresponding downstream dataset $\mathcal{D}_f^m, m \in \mathcal{M}$, which is used for task-specific fine-tuning of the downstream model θ_f . Our goal is to optimize the model performance on downstream tasks, formulated as $\max_{\theta_f} P(\theta_f(\mathcal{D}_f^m|\Theta))$. Here, Θ represents the initial weights from the pre-trained model θ_p , and P is the downstream performance metric, such as classification accuracy. Next, we define Multi-modal Transfer Gain to measure the impact of multi-modal pre-training on downstream tasks.

Definition 1 *Multi-modal Transfer Gain (Δ_{MTG}):* We denote the downstream model initialized by the multi-modal pre-

trained model as $\theta_f(\theta_p(\mathcal{D}_p))$, and the model initialized by the single-modal pre-trained model is denoted as $\theta_f(\theta_p(\mathcal{D}_p^m))$, $\mathcal{D}_p^m \in \mathcal{D}_p$. The Multi-modal Transfer Gain (Δ_{MTG}) is described as Eq. (1).

$$\Delta_{MTG} = P(\theta_f(\mathcal{D}_f^m|\theta_p(\mathcal{D}_p))) - P(\theta_f(\mathcal{D}_f^m|\theta_p(\mathcal{D}_p^m))). \quad (1)$$

We can further derive the definition of negative transfer:

Definition 2 *Multi-modal Negative Transfer:* The multi-modal negative transfer occurs if $\Delta_{MTG}(\theta_p(\mathcal{D}_p), \theta_p(\mathcal{D}_p^m)) < 0$. Otherwise, it is a multi-modal positive transfer.

The multi-modal transfer gain can reflect the degree to which

the knowledge of other modalities in the pre-trained foundation model affects the current downstream tasks of a specific modality. The multi-modal negative transfer indicates that the initialization weights from pre-training contain knowledge from other modalities that hinder performance on the specific downstream task. Next, we conduct a series of experiments to explore the presence and extent of negative transfer within the pre-trained and fine-tuned (PT-FT) process of the medical MFMs.

3.2 Effect of Negative Transfer

We pre-train four types of medical MFMs: (1) Single-modal pre-training. We pre-train the MAE [5] using data from a single modality. (2) Joint pre-training. We use UniMiSS [44] where modalities are trained together by universal embedding technique. (3) Sequential pre-training. We use MedCoSS [17] framework where different modalities are trained one by one. (4) Black-box model. We directly use the encoder of a well-pretrained MedSAM [26] to extract visual features. We select 10 distinct medical modalities to measure the extent of negative transfer for each method.

As shown in Fig. 1(b) and Fig. 1(c), only in a limited number of cases (e.g., Report, Pathology, Dermoscopy) does the MFM yield positive transfer gains for specific downstream tasks. More commonly, however, the weight initialization from multi-modal pre-training results in serious negative transfer gains. Unlike single-modal foundation models, which rely on a unique and cohesive feature space, MFMs generate multiple and diverse feature spaces that often conflict with each other, a phenomenon termed “modal collisions” [17], weakening the model’s representation learning for an individual modality. Negative transfer is not only a critical barrier to advancing MFMs, but also restricts their potential for clinical applications. These observations raise a natural and fundamental question: what factors influence the transferability of knowledge in MFMs? For this question, we further dive into the intricacies of the pre-training and fine-tuning process.

3.3 Exploring the Influence of Pre-training Epochs

The trajectory of pre-training is a critical factor in transfer learning, because models at different stages may encode features with varying levels of generality. Previous work has demonstrated that models initially capture low-frequency features (e.g., color and texture) that generalize well across domains, while high-frequency features learned later tend to specialize the model and reduce generalization ability [15]. Thus, we hypothesize that intermediate pre-training checkpoints can sometimes yield better downstream transfer

than fully converged models. To validate this hypothesis, we evaluated checkpoints sampled from ten stages of the pre-training process. Specifically, the 300-epoch joint pre-training trajectory was divided into ten equal intervals, and checkpoints were evaluated at these epochs. The results in Fig. 2 reveal two key insights: (1) transfer performance depends strongly on both modality and pre-training epoch, and no single epoch maximizes transferability for all. (2) Transfer gain does not change monotonically with the pre-training process. For instance, the 150th epoch shows superior transfer for the report modality. **The results imply that combining models from different pre-training epochs may help mitigate the negative transfer problem.**

3.4 Exploring the Influence of Fine-tuning Layers

Another key factor affecting transfer learning is downstream model fine-tuning. Previous studies have pointed out that foundation models request only partial fine-tuning to achieve an optimal trade-off between computing efficiency and transfer performance [5]. However, determining which specific layers to freeze remains an open question. For this reason, we analyze fine-tuning effects by dividing the image encoder layers into shallow (#1-#6), middle (#4-#10), deep (#7-#12), and a randomly 6-layer groups. This overlapping grouping strategy is employed to accommodate the continuous nature of feature transition within the network. Unlike rigid non-overlapping partitions, this approach ensures the transitional features that often span across the boundaries of shallow and deep layers are adequately included in the analysis, thereby providing a more comprehensive view of layer-wise transferability. We fine-tune these layers separately and freeze other parameters in the network. Fig. 3(a) reveals that partial fine-tuning achieves comparable results to full parameter fine-tuning, with the optimal layer groups differing by modality. For example, fine-tuning the middle and deep layers results in greater transfer gains for grayscale image modalities (e.g., CT, CXR). RGB image modalities (e.g. 2D pathology), in contrast, demonstrate superior performance with fine-tuning on randomly selected layers.

These observations raise a critical question: what differentiates the representational capacity of these layers for fine-tuning? Gradient norm analysis offers insight into the contribution of each layer’s parameters during fine-tuning. As shown in Fig. 3(b), middle and deep layers maintain active gradient norms throughout training, whereas shallow layers show reduced gradient activity from the middle stages of fine-tuning onward. Therefore, we believe that **a reasonable**

selection of the layers that contribute most to the network during fine-tuning may decrease the effect of negative transfer.

4 The MedRev Approach

In this section, we introduce an improved fine-tuning method for medical foundation models that reverses the negative transfer from multi-modal pre-training (MedRev).

4.1 Overview

The overall architecture of MedRev is shown in Fig. 4. The input $\mathbf{I} \in \{\mathbb{R}^{C \times L}, \mathbb{R}^{C \times H \times W}, \mathbb{R}^{C \times H \times W \times D}\}$ may consist of 1D, 2D or 3D data. C is the input channel, L is the input length, H, W and D are the input resolution. Depending on the input dimensionality, an appropriately selected tokenizer ($\{T^{1d}, T^{2d}, T^{3d}\}$) divides the data into non-overlapping patches $\mathbf{X} \in \mathbb{R}^{N \times P}$, where P represents the patch resolution and N is the total number of patches. These patches $\mathbf{X}$ are then projected onto an embedded space $\mathbf{x} \in \mathbb{R}^{K \times N}$ with a fixed dimension K . The embedded features are subsequently processed by the pre-trained ViT encoder, and they consist of a mixture of pre-trained experts (MoPE) and dynamic layer freezing (DLF) strategies. Finally, the encoder output is passed to a task-specific head, where either a multi-layer perceptron (MLP) or a feature pyramid network (FPN) is trained to generate the predicted classification or segmentation results.

4.2 Integrating Knowledge from Multiple Pre-Trained Experts

From our analysis in Section 3.3, a fixed pre-training checkpoint does not usually yield optimal results across all downstream tasks. Each modality may benefit from different pre-trained epochs. In light of the success of sparse mixture-of-experts (Sparse-MoE) [60], we propose a mixture of pre-trained experts that encourages the network to select requisite pre-trained knowledge based on downstream data. MoPE consists of a set of expert networks $E_1, E_2, \dots, E_n$, and a gate G . As shown in Fig. 4(a), the expert network represents a vision transformer encoder initialized from a certain pre-trained checkpoint. Given an input $\mathbf{x}$, the output of MoPE can be written as Eq. (2).

$$y = \sum_{i=1}^n G(x)^i E^i(x), \quad (2)$$

$$E^i(x) = \theta_p^i \cdot x, \quad i \in [1, \text{max epochs}],$$

where $G(x)$ and $E^i(x)$ are the outputs of the gate G and the i -th expert network. Next, the weight of each expert is

determined by the output of the gate, and we use a sparse Top-K gating based on the design of [60], i.e., the K highest contributions among all experts are selected to participate in the forward calculation and back-propagation [61, 62]. Eq. (3) describes the sparse gate formula.

$$G(x) = \text{Softmax}(\text{TopK}((x \cdot W_r), K)),$$

$$\text{TopK}(\alpha, k) = \begin{cases} \alpha_i & \text{if } \alpha_i \text{ in top } k \text{ elements of } \alpha \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where W_r is the learnable weight matrix of the gate. Notably, as the experiments in Section 3.3 demonstrate that early, middle, and late pre-training stages contribute differently in fine-tuning performance, we choose $n = 3$ experts and adopt the same uniform sampling principle to select evenly spaced checkpoints along the full pre-training trajectory so that each expert corresponds to a distinct stage of representation maturation. Besides, we use a Top-2 gate to accommodate the computing constraints of our GPU. This gating mechanism serves as an implicit and dynamic filter for modal conflicts. If the feature representation from a specific pre-training checkpoint conflicts with the current downstream task (i.e., it induces negative transfer and higher task loss), the back-propagation process drives the gate to assign lower weights to that expert, effectively suppressing the conflicting signal. Conversely, experts providing transferable and beneficial knowledge are prioritized. This allows MedRev to automatically identify and reducing the impact of modal conflicts during fine-tuning.

4.3 Adaptation to Parameter-Efficient Fine-Tuning

Given the versatility of our approach, we design two lightweight variants: parameter-efficient fine-tuning version (MedRev-PEFT) and linear probe version (MedRev-LP). For MedRev-PEFT, we adopt the low-rank approximation (LoRA) [10, 33], keeping total trainable parameters to only 1% of the full parameters. MedRev-LP offers an even lighter setup, training only the gating networks and a task-specific head. The overall trainable parameters are 0.5% of the full model. The architecture of these three variants is shown in Fig. 4(a).

4.4 Dynamic Selection of Learnable Layers

From Section 3.4, we know that manually setting these learnable layers through empirical experience is inefficient and difficult to optimize. It is better for the network to select the layers to be frozen based on the inputs. Therefore, we propose the dynamic layer freezing strategy. This is a simple and effective idea, which comes from the intuitive observation in our experiment: the gradient norm reflects, to some extent, the impact of a certain layer on the current representation

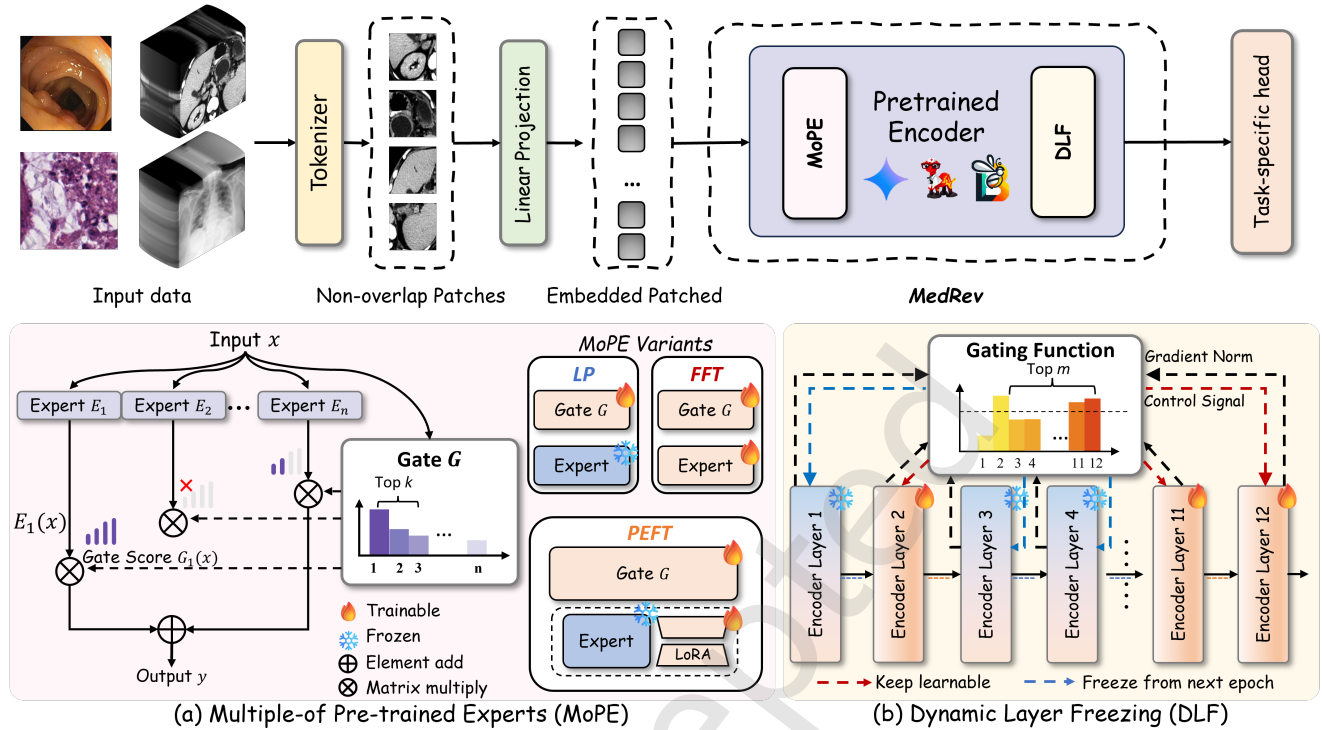


Fig. 4 Overview of the MedRev framework. (a) We replace each of the encoder blocks in the vision transformer with our proposed mixture of pre-trained experts (MoPE), where each expert represents an encoder block initialized from different pre-training checkpoints. The learnable gate is trained to select suitable experts and assign them weights. (b) The dynamic layer freezing is designed to determine which layers should be fine-tuned for certain modalities via a non-parameter gate function.

learning. In the early stage of fine-tuning, all layers are relatively active, so we keep 1/5 of the fine-tuning epochs for full parameter fine-tuning. Starting from 1/5 epoch, the contributions of different layers start to be clearly differentiated, and we compute the gradient norm for each layer after every epoch and feed it to a Top-K gate function, similar to Eq. (3). We choose layers with Top-6 gradient norm values to remain trainable and freeze the rest of the layers, as shown in Fig. 4(b).

4.5 Black-box Fine-tuning

It is worth noting that using MoPE requires conducting pre-training from scratch or having the developer provide model checkpoints from multiple epochs. However, most existing medical MFMs typically only release the last checkpoint for the community. Besides, due to privacy protection, we lack access to the pre-training data to update the model weights. To address this “black box” model problem, we use $n = 2$ experts, thereby adapting the method to dynamically aggregate from the ImageNet [63] initialization and the last checkpoint. This provides a practical two-expert approximation when only a single published checkpoint is available. We investigate the effectiveness of this approach in subsequent experiments.

5 Experiments

5.1 Dataset

All datasets used in this work are summarized in Table 1. This table reports the pre-training corpora covering ten modalities, the downstream fine-tuning datasets, their associated tasks, and the external generalization datasets. For each entry, this table provides sample counts, annotation types, literature references and download links.

5.2 Experimental Setup

5.2.1 Implementation Details

We adopt the ViT-B [64] pre-trained on ImageNet [63] as the default backbone. We choose ImageNet initialization as the default starting point because it provides a unified and widely used representation for all modalities, which helps isolate the effect of our proposed MoPE and DLF mechanisms from modality-specific pre-training biases. Following the input setting of MedCoSS [17], we set the length of 1D vectors to 112, the size of 2D patches to 224×224 , and the size of 3D patches to $16 \times 192 \times 192$. We apply random crop, resize, and flip as image augmentations. All models are trained with AdamW optimizer [65] at an initial learning rate of 1.5×10^{-4} ,

Table 1 Overview of the datasets used in this study.

Dataset	Modality	Task	#Train	#Val	#Test	Link
MIMIC-CXR-Report [42]	1D Report		227,323	-	-	
MIMIC-CXR-Image [42]	2D CXR		356,309	-	-	https://physionet.org/content/mimic-cxr-jpg/2.1.0/
DeepLesion [43]	3D CT		10,594	-	-	https://nihcc.app.box.com/v/DeepLesion
Brats [44]	3D MRI		2,040	-	-	https://www.kaggle.com/c/rsna-miccai-brain-tumor-radiogenomic-classification
NCH-100K [45]	2D Pathology	MLM/MIM	100,100	-	-	https://zenodo.org/records/1214456#.YwH6HOpBy5f
Eyepacs [46]	2D Fundus		35,126	-	-	https://zenodo.org/records/5793241
Ldopolypvideo [47]	2D Endoscopy		371,400	-	-	https://github.com/dashishi/LDPolypVideo-Benchmark
BUV [48]	2D Ultrasound		25,272	-	-	https://github.com/jhl-Det/CVA-Net
ISIC2024 [49]	2D Dermoscopy		81,722	-	-	https://www.kaggle.com/datasets/tomooiubushi/all-isic-data-20240629/
OCT 2017 [50]	2D OCT		84,484	-	-	https://data.mendeley.com/datasets/rschjbr9sj/2
PubMed-20k [51]	1D Report	Cls	180,040	30,212	30,135	https://github.com/Franck-Dernoncourt/pubmed-rct/tree/master
QaTa [52]	2D CXR	Seg	5,716	1,429	2,113	https://www.kaggle.com/datasets/aysenderli/qatacov19-dataset
RICORD [53]	3D CT	Cls	194	52	84	https://www.cancerimagingarchive.net/collection/midrc-ricord-1a/
VS [54]	3D MRI	Seg	193	-	49	https://www.cancerimagingarchive.net/collection/vestibular-schwannoma-seg/
NCH-7k [45]	2D Pathology	Cls	5,026	718	1,436	https://zenodo.org/records/1214456#.YwH6HOpBy5f
Messidor [54]	2D Fundus	Cls	1,223	174	351	https://www.adcis.net/en/third-party/messidor2/
PolypSet [55]	2D Endoscopy	Seg	1,450	-	798	https://github.com/DengPingFan/PraNet
BUSI [56]	2D Ultrasound	Seg	546	78	156	https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset
ISIC2017 [57]	2D Dermoscopy	Cls	2,000	150	600	https://challenge.isic-archive.com/data/#2017
OCT-C8 [58]	2D OCT	Cls	18,400	2,800	2,800	https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8
Private	2D Echocardiogram	Cls	9,072	1,296	2,592	https://github.com/yuncheng97/MedRev
Leukemia [59]	2D Microscopy	Cls	10,595	1,513	3,027	https://www.kaggle.com/datasets/andrewmvd/leukemia-classification

Table 2 Performance of LP-based (top), PEFT-based (middle), and FFT-based (bottom) tuning methods on classification datasets. The best results in each column are marked in bold. Δ_{MTG} is computed by ACC.

Tuning Method	Report		CT		Pathology		Fundus		OCT		Dermoscopy	
	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}
LP	82.2±0.3	0.1±0.11	83.0±0.8	-1.0±0.03	83.7±0.7	-1.1±0.11	74.7±0.6	-1.4±0.23	77.1±0.3	-2.2±0.18	79.8±0.4	-0.6±0.29
MedMerge [36]	82.5±0.3	0.3±0.18	83.8±0.4	-0.6±0.24	84.7±0.2	0.3±0.11	76.5±0.5	0.2±0.22	78.1±0.6	-1.4±0.51	80.5±0.6	0.5±0.28
MedRev-LP	82.7±0.2	0.7±0.25	84.0±0.7	-0.2±0.13	85.4±0.4	0.8±0.12	77.1±0.6	0.6±0.07	79.6±0.4	-0.2±0.16	81.2±0.2	1.1±0.18
LoRA [33]	83.6±0.3	0.9±0.17	86.8±0.6	-0.2±0.25	92.9±0.8	0.4±0.21	82.5±0.5	-0.2±0.33	86.0±0.3	-0.2±0.18	84.7±0.2	0.7±0.27
SSF [34]	83.1±0.2	0.3±0.19	86.0±0.5	-0.8±0.24	92.0±1.1	-0.5±0.09	81.4±0.5	-0.9±0.04	84.4±0.1	-1.1±0.02	83.6±0.5	-0.8±0.29
Adapter [9]	82.8±0.6	0.5±0.23	86.3±0.7	0.2±0.31	92.6±0.4	0.3±0.20	81.9±0.9	-0.4±0.15	85.5±0.2	-0.2±0.08	83.5±0.5	-0.3±0.24
TPP [35]	83.9±0.4	1.1±0.09	87.1±0.2	0.1±0.20	92.3±0.1	-0.1±0.05	82.8±0.7	0.3±0.06	86.6±0.3	0.5±0.23	84.5±0.1	0.7±0.05
MedRev-PEFT	84.4±0.2	1.5±0.12	87.6±0.2	0.6±0.05	93.4±0.3	1.0±0.7	83.4±0.4	0.9±0.22	87.5±0.4	1.3±0.09	85.7±0.2	1.7±0.11
FFT	83.7±0.5	0.7±0.15	86.5±0.3	-1.7±0.29	91.6±0.1	0.0±0.2	80.2±0.5	-1.1±0.16	86.7±0.3	-1.0±0.33	84.8±0.6	0.5±0.13
BSS [12]	84.0±0.2	1.1±0.15	87.3±0.6	-0.5±0.21	91.9±1.0	0.2±0.37	80.8±0.2	-0.9±0.11	87.1±0.2	-0.5±0.14	85.0±0.2	0.6±0.07
Wang et al. [16]	83.5±0.3	0.9±0.20	87.1±0.4	-0.1±0.09	91.3±0.8	-0.2±0.33	80.6±0.3	0.1±0.22	86.9±0.6	-0.3±0.27	83.7±0.5	-0.1±0.04
PCGrad [32]	83.7±0.7	0.8±0.18	88.2±0.3	0.2±0.11	92.1±0.3	0.5±0.28	80.5±0.4	-0.6±0.08	87.0±0.2	0.2±0.10	85.2±0.3	0.6±0.09
ConceptTune [37]	84.2±0.5	1.3±0.03	87.6±0.5	-0.3±0.06	92.0±0.7	-0.1±0.2	81.2±0.3	-0.4±0.13	86.4±0.5	-0.6±0.05	85.5±0.3	1.3±0.10
SpotTune [72]	83.9±0.5	0.8±0.14	88.0±0.4	0.2±0.08	91.5±0.1	0.6±0.06	80.4±0.5	0.1±0.05	87.8±0.3	0.2±0.05	84.9±0.2	0.5±0.02
Yang et al. [39]	84.4±0.8	1.1±0.23	87.9±0.7	0.1±0.26	91.7±0.9	0.1±0.41	81.0±0.7	0.2±0.18	87.3±0.5	0.3±0.17	85.6±0.4	0.9±0.16
MedRev-FFT	84.9±0.7	2.0±0.06	89.0±0.6	0.9±0.15	92.8±0.3	1.4±0.25	81.7±0.4	0.5±0.09	88.8±0.3	1.0±0.14	86.0±0.2	1.6±0.30

following a constant learning rate schedule with a warmup phase. Pre-training takes 300 epochs on eight NVIDIA V100 GPUs with a batch size of 512, while fine-tuning takes 64 epochs on one NVIDIA V100 GPU.

5.2.2 Comparison Methods and Evaluation Metrics

For downstream tasks, our method is compared with 13 popular fine-tuning methods grouped into three categories: linear probe (LP), parameter-efficient fine-tuning (PEFT), and full fine-tuning (FFT). (1) For LP, we compare vanilla LP and MedMerge [36]. (2) For PEFT, we compare LoRA [33], SSF [34], Adapter [9], and TPP [35]. (3) For FFT, we compare BSS [12], Wang et al. [16], PCGrad [32], ConceptTune [37], SpotTune [66], and Yang et al. [39]. All task-specific baselines were re-implemented using the publicly released code and recommended default hyperparameters. Following MedCoss [17], the fine-tuning performance was evaluated by predictive accuracy (ACC) and the area under the receiver

operator curve (AUC) for image classification task. The dice similarity coefficient (DSC) and the 95% Hausdorff distance (HD) were employed to measure the image segmentation task. In addition, we use the multi-modal transfer gain (Δ_{MTG}) to evaluate the model's capability of mitigating negative transfer. To ensure the robustness of our conclusions, we report the mean and standard deviation across three random seeds and perform paired t-tests [67–74] on the same test splits against the main baseline of each task, with significance reported at the two-sided $p < 0.05$ level unless otherwise noted.

5.3 Quantitative Results

5.3.1 Results on Classification Datasets

We evaluate the classification performance across six medical datasets. Table 2 provides a comprehensive comparison of different tuning strategies. While standard fine-tuning approaches such as LP, FFT, and several PEFT methods achieve

Table 3 Performance of LP-based (top), PEFT-based (middle), and FFT-based (bottom) tuning methods on segmentation datasets. The best results in each column are marked in bold. Δ_{MTG} is computed by DSC.

Tuning Method	MRI		CXR		Ultrasound		Endoscopy	
	DSC(%)	Δ_{MTG}	DSC(%)	Δ_{MTG}	DSC(%)	Δ_{MTG}	DSC(%)	Δ_{MTG}
LP	82.1±0.5	-1.3±0.11	72.7±0.4	-2.1±0.06	72.4±0.2	-0.8±0.08	75.6±0.5	-1.3±0.04
MedMerge [36]	82.7±0.3	-0.3±0.19	74.1±0.1	-1.4±0.14	72.5±0.6	-0.5±0.08	76.3±0.3	-0.8±0.13
MedRev-LP	83.5±0.3	-0.1±0.03	74.8±0.2	-0.3±0.11	72.8±0.2	-0.3±0.15	77.6±0.4	0.5±0.09
LoRA [33]	87.0±0.6	-0.8±0.03	80.2±0.9	0.4±0.24	81.0±0.5	0.6±0.17	83.1±0.7	-0.9±0.07
SSF [34]	86.2±0.3	-1.3±0.13	79.2±0.2	-0.2±0.08	78.5±0.1	-1.2±0.11	83.0±0.9	-1.1±0.25
Adapter [9]	86.5±0.9	-0.9±0.19	79.4±1.0	0.2±0.09	79.4±0.3	-0.4±0.10	82.7±0.5	-0.7±0.04
TPP [35]	87.2±0.8	-0.4±0.12	79.6±0.7	0.0±0.07	80.3±0.6	0.2±0.2	83.4±0.8	-0.5±0.06
MedRev-PEFT	87.0±0.5	-0.4±0.23	80.7±0.4	1.1±0.12	80.3±0.4	0.5±0.22	84.2±0.3	0.4±0.05
FFT	87.9±0.5	-1.9±0.05	78.6±0.6	-0.8±0.17	79.7±0.3	-1.3±0.28	82.7±0.8	-1.9±0.11
BSS [12]	88.5±0.4	-1.4±0.22	79.4±0.8	-0.2±0.08	79.5±0.9	-1.2±0.05	83.0±0.7	-1.1±0.20
Wang et al. [16]	87.7±0.5	-0.7±0.31	78.3±0.3	0.1±0.12	79.3±0.7	-0.1±0.13	83.5±0.4	-0.6±0.13
PCGrad [32]	88.2±0.3	-1.1±0.27	78.8±0.5	-0.2±0.15	80.3±0.4	0.1±0.22	83.0±0.6	-0.6±0.22
ConceptTune [37]	88.0±0.2	-1.4±0.33	78.4±0.5	-0.7±0.18	79.8±0.3	-0.9±0.29	82.9±1.0	-0.7±0.05
SpotTune [72]	88.6±0.2	-1.0±0.12	79.2±0.3	0.0±0.06	80.1±0.5	-0.5±0.21	83.2±0.9	-0.8±0.18
Yang et al. [39]	88.4±0.3	-0.7±0.18	79.4±0.7	0.3±0.19	79.8±0.5	-0.6±0.18	83.2±0.8	-0.8±0.14
MedRev-FFT	89.2±0.1	-0.3±0.08	80.2±0.3	0.9±0.11	80.8±0.1	-0.3±0.12	84.1±0.3	-0.1±0.07

Table 4 Quantitative comparison with SOTA classification methods. The best results in each column are marked in bold.

Method	CT		Pathology	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)
ViT [64]	86.7±0.2	85.8±0.3	90.2±0.8	98.7±0.3
MedViT [75]	87.2±0.7	88.2±1.0	92.7±0.9	98.9±0.4
MedFormer [76]	88.6±1.1	87.7±0.7	92.7±0.5	99.7±0.1
Incarnag [77]	88.1±0.9	88.4±1.6	93.0±0.2	99.1±0.3
MedRev-FFT	89.0±0.6	89.1±0.5	93.4±0.3	99.6±0.3

Table 5 Quantitative comparison with SOTA segmentation methods. The best results in each column are marked in bold.

Method	CXR		Endoscopy	
	DSC(%)	HD	DSC(%)	HD
UNet [78]	76.6±1.2	32.3±2.6	80.3±0.4	39.5±1.3
nnUNet [79]	81.3±0.7	27.5±0.7	84.7±0.8	36.7±0.2
TransUNet [80]	79.5±1.0	30.1±0.9	83.0±1.1	37.1±0.5
Swin-UNet [81]	80.2±0.9	28.9±1.1	82.9±0.7	37.8±1.0
MedRev-FFT	80.7±0.7	28.4±1.5	84.2±0.3	35.7±2.2

competitive accuracy in certain domains, many exhibit inconsistent or even negative transfer gains. For instance, ConceptTune [37] shows negative transfer Δ_{MTG} on CT, Pathology, Fundus, and OCT. This suggests that naïve pre-training adaptation still leads to harmful updates for some target domains. **In contrast, MedRev consistently outperforms prior techniques across all datasets.** MedRev-FFT, with fully trainable parameters, achieves the best ACC and Δ_{MTG} on Report, CT, and Dermoscopy datasets, all yielding positive Δ_{MTG} compared with single-modal pre-training, and statistically significantly better ($p < 0.05$) than FFT, BSS [12], SpotTune [66] on those cases. Beyond extensive fine-tuning, our lightweight version,

MedRev-PEFT, achieves a superior trade-off, which matches the accuracy of stronger FFT baselines while maintaining parameter efficiency on Pathology and Fundus.

5.3.2 Results on Segmentation Datasets

Subsequently, we evaluate these methods on four segmentation datasets. Image segmentation is a more challenging task as it requires dense pixel-level predictions, which are more prone to overfitting. As shown in Table 3, previous fine-tuning approaches often involve heavy parameterization and complex frameworks, which yield suboptimal results compared to lighter models. For instance, ConceptTune [37] achieves 78.4% for DSC, which is less than Adapter’s [9] 79.4%. And its Δ_{MTG} is -0.7, which further lags behind 0.2. **MedRev’s flexible design allows it to be integrated with parameter-efficient fine-tuning schemes, reducing the number of trainable parameters and the risk of overfitting..** Notably, MedRev-PEFT statistically significantly outperforms ($p < 0.01$) TPP [35] and LoRA [33] on CXR and Ultrasound dataset. Moreover, MedRev-PEFT approaches offer a practical advantage for clinical applications, achieving comparable performance with only 1% of the learnable parameters needed by FFT-based methods, which is more suitable for resource-limited scenarios like GPU constraints or fast deployment.

5.3.3 Compared With SOTA Methods

To evaluate the performance of MedRev compared with other task-specific state-of-the-art (SOTA) methods, we conduct experiments to compare on both classification and segmentation tasks. Table 4 reports the results for classification tasks on CT

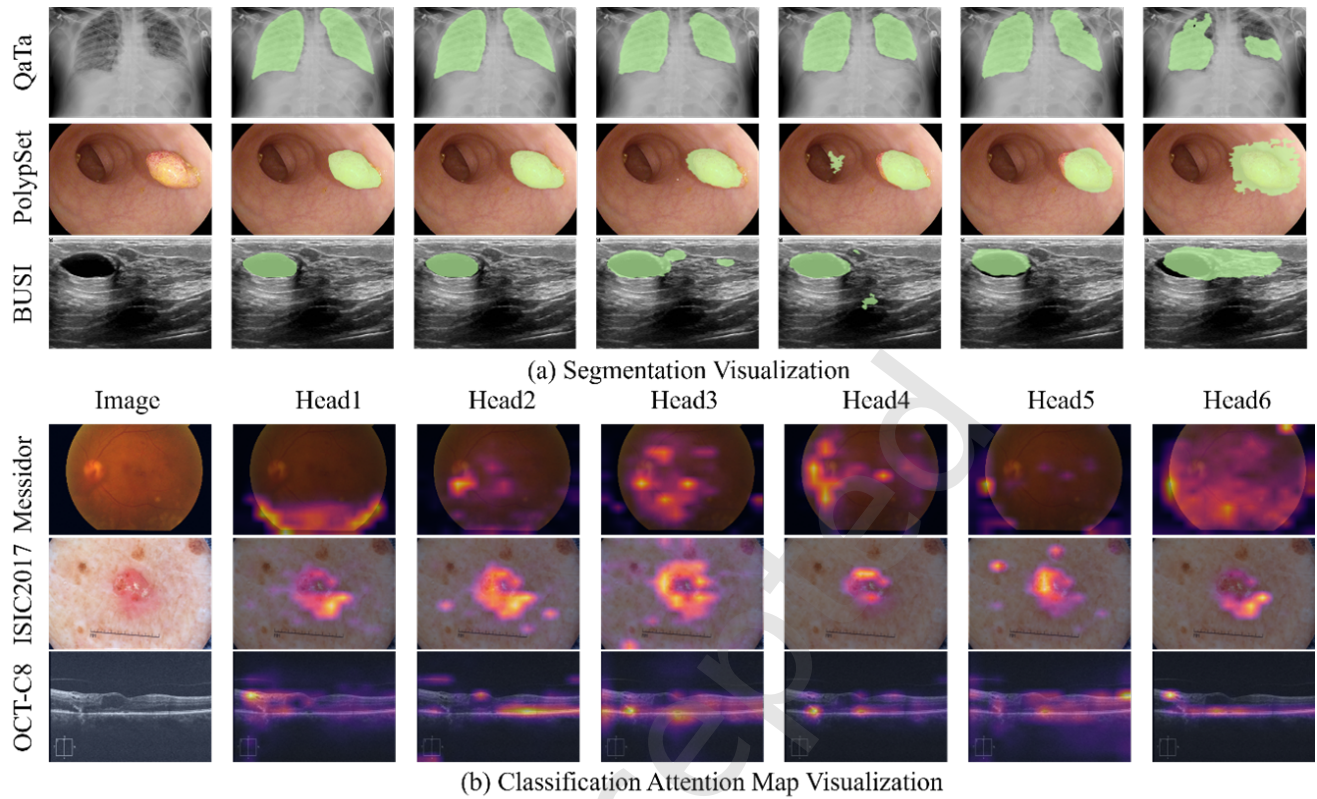


Fig. 5 Qualitative visualization. (a) Segmentation results: Predicted masks are displayed as translucent green overlays. (b) Classification attention maps: attention maps from six transformer heads. Warmer colors denote higher attention weights.

Table 6 Results on the unseen modality generalization capability. The best results in each column are marked in bold.

Method	Private (Echocardiogram)		Leukemia (Microscopic)	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)
FFT	82.8±0.5	91.0±0.3	82.3±0.9	90.8±0.5
BSS	84.8±0.4	93.0±0.8	84.1±0.6	92.5±0.7
ConceptTune	86.7±0.5	92.4±1.1	86.6±0.3	92.9±0.9
SpotTune	85.4±0.1	93.4±0.2	87.1±0.3	92.4±0.4
MedRev-FFT	87.2±0.4	95.6±0.6	88.3±0.8	93.7±0.3

and pathology datasets. All baseline results were reproduced on our datasets using the official implementations and recommended default settings. These methods, however, are limited by the absence of domain-specific medical knowledge and anatomical context. In contrast, MedRev was pre-trained on multi-modal datasets, allowing it to leverage knowledge from multiple modalities to increase task-specific performance. Specifically, MedRev achieves the best ACC (89%) and AUC (89.1%) on the CT dataset, statistically significantly better ($p < 0.05$) than previous SOTA methods. Also, it achieves 93.4% ACC and 99.6% AUC on pathology dataset, respectively. Table 5 presents the segmentation results on CXR and endoscopy datasets. While nnUNet [87] achieves the best scores, owing to its robust data preprocessing and dynamic configuration, MedRev consistently attains the second-best

Table 7 Computational cost comparison on the CT dataset. MedRev uses a shared backbone with three experts and Top-2 sparse gating. All measurements were obtained using the same NVIDIA V100 GPU, batch size, and numerical-precision setting.

Method	Trainable Parameters (M)	Peak Memory (GB)	Training Time (h)	Inference Time (ms/img)
LP	0.3	3.5	0.8	4.2
LoRA	0.9	4.8	1.5	7.8
FFT	89	9.4	2.5	12.1
MedRev-LP	0.5	8.5	1	5.6
MedRev-PEFT	1.2	10.5	1.6	9.1
MedRev-FFT	108	18.2	2.2	13.8

scores across two datasets, further highlighting its effectiveness in medical image segmentation tasks. The comparison against ImageNet-initialized backbones is meant to reflect a common and fair fine-tuning baseline; we therefore interpret these results as evidence that MedRev improves adaptation under a standard initialization regime

To further examine whether MedRev depends on ImageNet initialization, we compared ImageNet-initialized and medical-domain-initialized backbones under both FFT and MedRev settings. As shown in Table 9, medical-domain initialization provides a small absolute improvement over ImageNet initialization, while MedRev consistently improves both initialization settings. This result supports our use of ImageNet initialization as a controlled and reproducible baseline

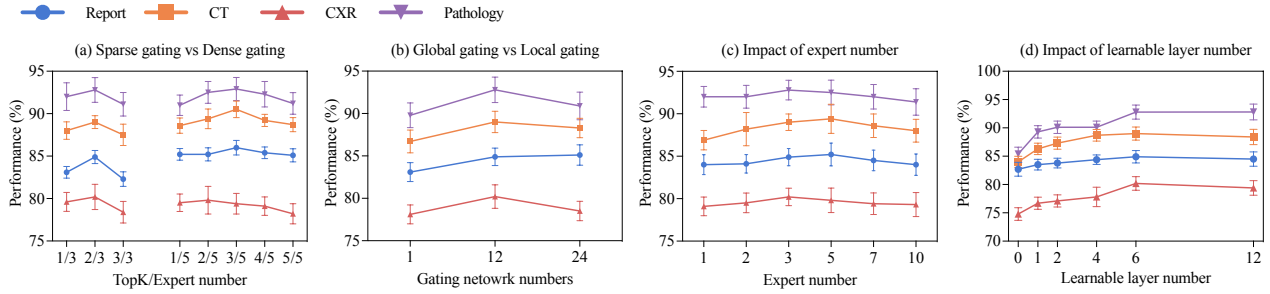


Fig. 6 Sensitivity analysis and ablation studies of the MedRev framework. (a) Performance variations under different gating strategies (sparse vs. dense) and *Top-K* selection. (b) Impact of the gating network count on multi-task coordination. (c-d) Sensitivity analysis of the number of experts and learnable layers. Markers indicating the mean and error bars representing standard deviation.

Table 8 The fully fine-tuning performance of different black-box pre-trained model on four selected modality datasets.

Pre-trained Model	Tuning Method	CT	MRI	CXR	Path.
		ACC(%)	DSC(%)	DSC(%)	ACC(%)
BiomedGPT [27]	FFT	88.2	89.7	82.1	93.0
	SpotTune	88.4	90.3	82.7	92.7
	MedRev-FFT	88.7	91.0	82.9	93.6
MedSAM [26]	FFT	87.7	88.6	79.7	91.8
	SpotTune	88.4	88.9	80.1	92.2
	MedRev-FFT	88.4	90.1	80.8	92.4
MedCoSS [17]	FFT	86.5	88.2	80.4	91.6
	SpotTune	86.8	88.2	80.8	91.9
	MedRev-FFT	87.4	88.8	81.6	92.9
LVM-Med [28]	FFT	88.0	89.3	82.5	92.7
	SpotTune	88.7	89.4	82.9	93.5
	MedRev-FFT	89.2	90.6	84.1	93.8

Table 9 Initialization robustness analysis for ImageNet initialization versus medical-domain initialization.

Task	Metric	ImageNet Init + FFT	Medical Init + FFT	ImageNet Init + MedRev	Medical Init + MedRev
CT	ACC(%)	86.5±0.6	87.2±0.5	89.0±0.6	89.3±0.4
CT	AUC(%)	88.0±0.5	88.4±0.4	89.1±0.5	89.5±0.4
Pathology	ACC(%)	91.6±0.4	92.3±0.3	93.4±0.3	93.7±0.3
Pathology	AUC(%)	99.1±0.2	99.3±0.2	99.6±0.3	99.7±0.2
CXR	DSC(%)	78.6±0.7	79.9±0.6	80.7±0.7	81.0±0.5
CXR	HD	29.7±1.1	29.3±1.0	28.4±1.5	28.1±1.2
Endoscopy	DSC(%)	82.7±0.6	83.2±0.5	84.2±0.3	84.5±0.4
Endoscopy	HD	37.0±0.6	36.8±0.5	35.7±2.2	35.3±1.7

rather than a method-specific requirement.

5.3.4 Assessing Domain Generalization Ability

Downstream fine-tune benefits from the modalities seen during upstream pre-training, thus facilitating transfer and adaptation. A robust medical model should not only effectively summarize the knowledge from seen modalities, but also generalize to unseen ones. To measure such generalization ability, we conduct experiments on two datasets, where the modality was not present in the pre-training data. As shown in Table 6, the private dataset consists of echocardiogram images, which are similar to ultrasound but have distinct textures and anatomical distributions. The Leukemia dataset comprises microscopic images, which differ from all other modalities. Our MedRev, by leveraging the rich pre-training knowledge and appropriate updated layers, consistently out-

performs previous fine-tuning methods, achieving the best performance on both datasets.

5.4 Visualization Results

Fig. 5 provides qualitative examples that reveal systematic behaviours that explain MedRev’s empirical gains. From the segmentation overlays in Fig. 5(a), **we can observe that MedRev recovers small and fragmented lesions with fewer false negative than baselines**, indicating stronger localization of weak signals. The multi-head attention visualizations in Fig. 5(b) show that this head-wise complementarity appears to reduce negative transfer across tasks by enabling diverse feature pathways, a behavior consistent with the model’s improved generalization.

5.4.1 Quantitative Analysis of Modality Conflicts

To provide a more direct characterization of inter-modality conflicts, we further quantified the modality-wise gradient cosine similarity during downstream fine-tuning. As shown in Fig. 7, the diagonal entries remain positive, whereas many cross-modality entries become negative, especially between text-like report features and image modalities, and between grayscale volumetric modalities and RGB-like modalities. The conflict ratios exceed 55% for all modalities. These observations provide direct quantitative evidence that negative multi-modal transfer is associated with conflicting optimization directions across modalities.

5.5 Discussions

5.5.1 Evaluation on Black-box Model Challenge

In Section 4.5, we discussed scenarios where only a single checkpoint of the foundation model is available, without access to the upstream pre-training data. In these cases, MoPE must be simplified to a two-expert fusion. We evaluated this approach on four well-established medical MFMs:

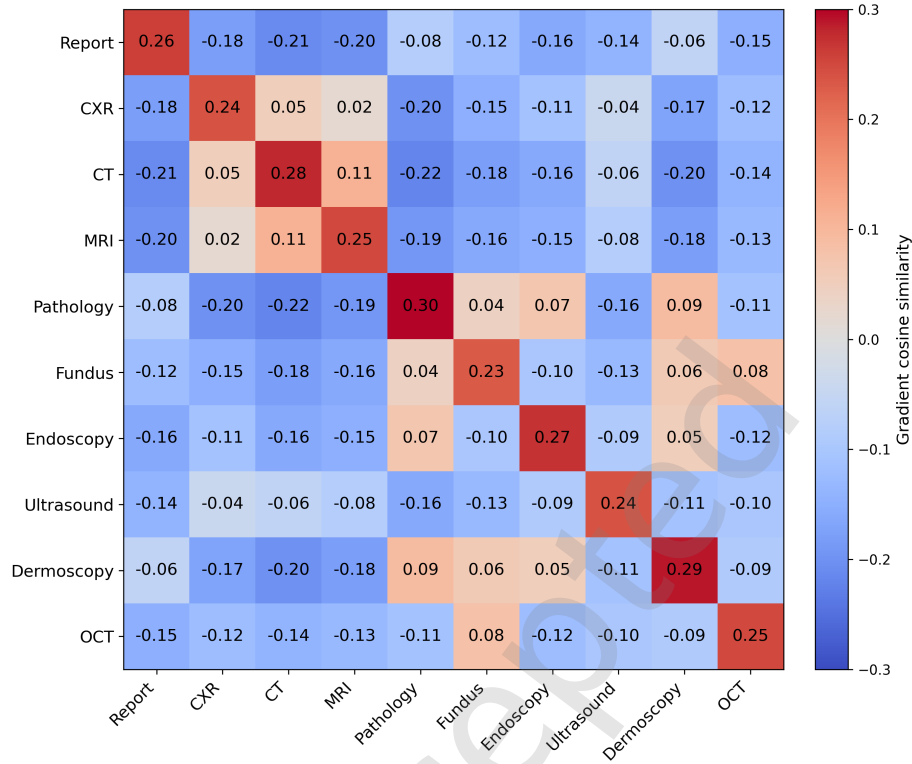


Fig. 7 Modality-wise gradient cosine similarity for quantifying inter-modality conflicts. Positive values indicate compatible optimization directions, while negative values indicate conflicting gradients.

BiomedGPT [27], MedSAM [26], MedCoSS [17], and LVM-Med [28]. Experiments were conducted on the four downstream datasets. As shown in Table 8, the two-expert fusion strategy still exhibits advantages over fine-tuning based on a single pre-trained weight. Combined with the results in Table 2 and 3, we further speculate that including intermediate epoch checkpoints could yield additional performance gains.

We further compared historical multi-modal checkpoint experts with a single-modal expert initialization alternative. As shown in Table 10, historical multi-modal checkpoints provide consistently higher performance than single-modal expert initialization across representative classification and segmentation tasks. This supports our design choice that experts from the same multi-modal training trajectory preserve architectural and optimization consistency while still offering diverse stages of representation maturity.

5.5.2 Sparse Gating vs Dense Gating

As shown in Fig. 6(a), MedRev with Top-2 sparse gating outperforms other approaches across four tasks. The sparsity helps mitigate optimization interference between tasks by focusing on the most balanced combination of experts. Denser gating, such as Top-3 or higher, appears to introduce excessive noise from non-essential experts, decreasing performance,

Table 10 Comparison between historical multi-modal checkpoint experts and single-modal expert initialization.

Task	Metric	Single-modal Expert MoPE	Historical MFM Checkpoint MoPE	MedRev
Report	ACC(%)	84.0±0.3	84.6±0.4	84.9±0.7
Report	Δ_{MTG}	1.3±0.08	1.8±0.07	2.0±0.06
CT	ACC(%)	88.1±0.5	88.8±0.4	89.0±0.6
CT	Δ_{MTG}	0.4±0.12	0.8±0.10	0.9±0.15
Pathology	ACC(%)	92.2±0.4	92.4±0.3	92.8±0.3
Pathology	Δ_{MTG}	0.8±0.10	1.1±0.15	1.4±0.25
Fundus	ACC(%)	81.9±0.6	82.7±0.5	83.0±0.4
Fundus	Δ_{MTG}	0.3±0.11	0.9±0.08	1.0±0.09
CXR	DSC(%)	79.5±0.6	79.8±0.5	80.2±0.3
CXR	Δ_{MTG}	0.2±0.09	0.8±0.07	0.9±0.11
Endoscopy	DSC(%)	83.4±0.5	83.8±0.4	84.1±0.3
Endoscopy	Δ_{MTG}	-0.8±0.10	-0.5±0.08	-0.1±0.07

while overly sparse gating (Top-1) risks omitting essential expert knowledge, leading to suboptimal results.

5.5.3 Global Gating vs Local Gating

MedRev allows for multi-granularity gating options. Global gating means that the expert network is the entire pre-trained vision transformer encoder with only one gating network. Local gating means that the expert network is the vision transformer encoder block (12 gates), or the self-attention module and FFN module (24 gates). As shown in Fig. 6(b), this local gating strategy allows MedRev to adaptively select various experts for different parts of the network, which may be

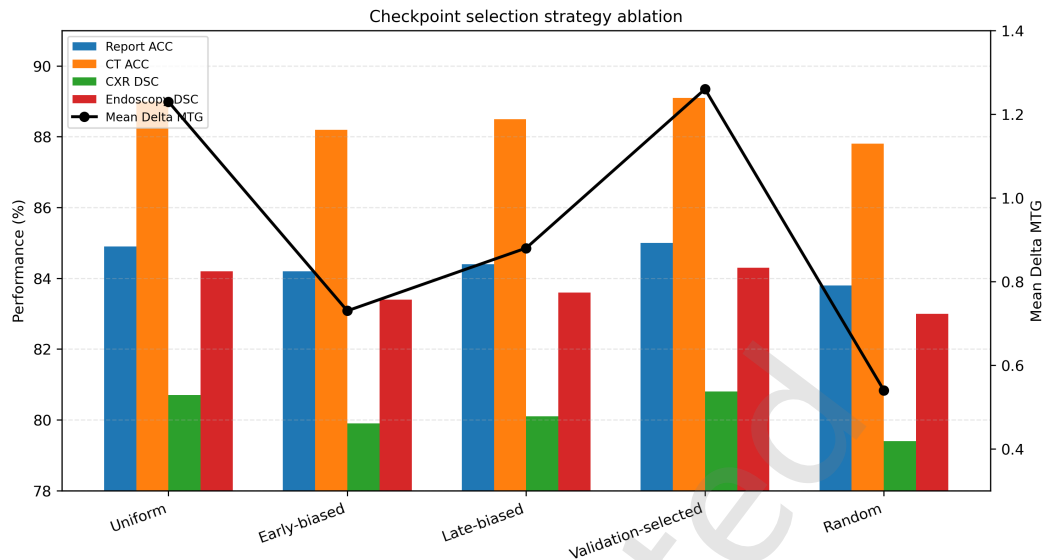


Fig. 8 Checkpoint selection strategy ablation. Uniform checkpoint sampling achieves a strong trade-off between performance, transfer gain and reproducibility.

Table 11 Ablation studies of the proposed methods.

Method	Report		CT	
	ACC(%)	Δ_{MTG}	ACC(%)	Δ_{MTG}
w/o MoPE & DLF	82.8 $\pm$ 0.2	0.7 $\pm$ 0.13	86.5 $\pm$ 0.6	-0.7 $\pm$ 0.15
w/o MoPE	84.0 $\pm$ 0.1	0.5 $\pm$ 0.08	86.9 $\pm$ 0.3	0.2 $\pm$ 0.08
w/o DLF	84.5 $\pm$ 0.7	1.1 $\pm$ 0.20	88.4 $\pm$ 0.7	0.5 $\pm$ 0.23
w Model Averaging	83.7 $\pm$ 0.1	0.9 $\pm$ 0.12	86.8 $\pm$ 0.5	-0.2 $\pm$ 0.31
w Switch MoPE [82]	85.2 $\pm$ 0.4	1.5 $\pm$ 0.06	88.1 $\pm$ 0.7	0.4 $\pm$ 0.03
w MoFE [83]	84.4 $\pm$ 0.5	1.0 $\pm$ 0.04	88.3 $\pm$ 0.3	0.2 $\pm$ 0.05
MedRev	84.9 $\pm$ 0.7	2.0 $\pm$ 0.06	89.0 $\pm$ 0.6	0.9 $\pm$ 0.15

particularly beneficial for handling diverse downstream tasks with varied feature requirements. The success of local gating demonstrates that fine-grained control over expert selection enables task-specific knowledge integration, increasing the model’s ability to generalize across multiple modalities.

5.5.4 Computational Efficiency Analysis

To evaluate the computational efficiency of MedRev, we compared its cost with LP, LoRA, and FFT on the CT dataset. Table 7 reports trainable parameters, peak training memory, training time, and inference latency measured under identical hardware and training settings. The default MedRev configuration uses a shared backbone, three experts, and Top-2 sparse gating. Compared with standard FFT, MedRev-FFT increases the number of trainable parameters from 89M to 108M (21.3%) and peak training memory from 9.4 GB to 18.2 GB (93.6%). The disproportionate memory increase indicates that the overhead should not be attributed to the gate parameters alone; rather, it is consistent with the additional trainable states and the activations and gradients retained for the two selected expert paths during back-propagation. Because only

two experts are executed for each sample, inference latency increases more moderately, from 12.1 to 13.8 ms/image (14.0%). MedRev-FFT requires 2.2 h for training compared with 2.5 h for standard FFT. For resource-constrained settings, the validated MedRev-LP and MedRev-PEFT variants reduce the trainable parameter count to 0.5M and 1.2M and peak training memory to 8.5 GB and 10.5 GB, respectively.

5.6 Ablation Studies and Hyperparameter Analysis

5.6.1 Checkpoint Selection Strategy

To further justify the checkpoint selection strategy, we compared uniformly sampled checkpoints with early-biased, late-biased, validation-selected and random checkpoint selections. As shown in Fig. 8, validation-selected checkpoints slightly improve the mean transfer gain but require task-specific tuning, whereas uniformly sampled checkpoints achieve nearly identical performance without using downstream validation information for checkpoint selection. Random checkpoints perform worst, indicating that checkpoint coverage across the training trajectory is important. These results support our choice of uniformly sampled historical checkpoints as a fair and reproducible strategy.

5.6.2 Effectiveness of Modules

To validate the effectiveness of the proposed methods, we conducted ablation studies across four medical imaging modalities. As shown in the first block of Table 11, removing both MoPE and DLF results in significant performance degradation, particularly in CT (ACC -2.5%) and CXR (DSC -1.6%). Introducing either DLF or MoPE individually led

to notable improvements, highlighting their critical roles in multi-modal transfer learning. Combining MoPE with DLF (Ours) achieved the best performance, demonstrating the synergistic effect of adaptive expert fusion and parameter optimization. These ablations support the narrative that both checkpoint diversity and layer selection are necessary to reverse negative transfer.

5.6.3 Impact on the MoE Strategy

Furthermore, the proposed MoPE serves as a plug-and-play strategy that can be integrated with existing ensemble approaches or MoE frameworks. As shown in the second block of Table 8, traditional ensemble methods exhibited relatively limited improvements, likely due to their simplistic aggregation mechanisms, which are insufficient for complex multi-modal tasks. In contrast, more advanced designs such as Switch-MoE [82] achieve comparable performance to MedRev, demonstrating the scalability and generalizability of the proposed method.

5.6.4 Impact on Numbers of Experts

The overhead of MoPE scales differently for storage and per-sample computation. The backbone is shared and stored only once, so increasing the total number of experts, N , mainly increases the storage required for expert-specific parameters and their training states. By contrast, Top-2 gating activates only two experts for each input. Thus, with $K = 2$ fixed, the main per-sample expert computation is determined by the two active experts rather than by the total number of stored experts. As shown in Fig. 6(c), increasing the number of experts from three to five yields less than a two-percentage-point gain, while performance decreases beyond five experts, supporting three experts as a practical default.

For resource-constrained settings, the validated MedRev-LP and MedRev-PEFT variants provide immediate lower-memory alternatives. Further potential optimizations include reducing N or K when the performance trade-off is acceptable, using low-rank expert-specific modules, mixed-precision training, or activation checkpointing, and pruning infrequently selected experts after training to reduce model storage.

5.6.5 Impact on Numbers of Learnable Layers

We further examined the effect of varying the number of learnable layers in DLF. As shown in Fig. 6(d), fine-tuning only a single transformer layer already boosts accuracy compared to the linear probing from 85.4% to 89.3% on the pathology dataset. Fine-tuning four to six layers achieves performance comparable to full fine-tuning, while the latter can suffer slight degradation from overfitting. Overall,

fine-tuning approximately half of the layers offers the best trade-off between efficiency and performance.

6 Conclusion

In this paper, we present a systematic exploration into the occurrence and causes of negative multi-modal transfer. Extensive experiments demonstrate that this phenomenon is widespread across modalities. Our analysis further reveals that transferability is closely linked to both the multi-modal knowledge captured at different checkpoints during pre-training and the selection of learnable layers during fine-tuning. Building on these insights, we propose MedRev, a simple yet effective fine-tuning paradigm developed to mitigate negative transfer. MedRev integrates a mixture of pre-trained experts with a dynamic layer freezing strategy, allowing the model to leverage appropriate multi-modal priors while adaptively updating task-relevant layers. Evaluations on ten downstream datasets confirm the consistent superiority of MedRev.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62372316), Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2024ZD0532900), and the Sichuan Science and Technology Program Key Project (Nos. 2024YFHZ0091, 2025YFHZ0066).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

- [1] W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang, A comprehensive survey of foundation models in medicine, *IEEE Reviews in Biomedical Engineering*, 2025.
- [2] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, Foundation models for generalist medical artificial intelligence, *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.
- [3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *International conference on machine learning*. Pmlr, 2020, pp. 1597–1607.
- [4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, Dinov2: Learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193*, 2023.

- [5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, Masked autoencoders are scalable vision learners, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.
- [6] Y. Jiang, C.-M. Feng, J. Ren, J. Wei, Z. Zhang, Y. Hu, Y. Liu, R. Sun, X. Tang, J. Du *et al.*, From pretraining to privacy: federated ultrasound foundation model with self-supervised learning, *npj Digital Medicine*, vol. 8, no. 1, p. 714, 2025.
- [7] S. Kornblith, J. Shlens, and Q. V. Le, Do better imagenet models transfer better? in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2661–2671.
- [8] Z. Jiang, D. Cheng, Z. Qin, J. Gao, Q. Lao, A. B. Ismoilovich, U. Gayrat, Y. Elyorbek, B. Habibullo, D. Tang *et al.*, Tv-sam: increasing zero-shot segmentation performance on multimodal medical images using gpt-4 generated descriptive prompts without human annotation, *Big Data Mining and Analytics*, vol. 7, no. 4, pp. 1199–1211, 2024.
- [9] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, Parameter-efficient transfer learning for nlp, in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.
- [10] Y. You, K. Tan, Z. Jiang, and L. Zhang, Developing a predictive platform for salmonella antimicrobial resistance based on a large language model and quantum computing, *Engineering*, vol. 48, pp. 174–184, 2025.
- [11] J. Gao, Q. Lao, Q. Kang, P. Liu, C. Du, K. Li, and L. Zhang, Boosting your context by dual similarity checkup for in-context learning medical image segmentation, *IEEE Transactions on Medical Imaging*, vol. 44, no. 1, pp. 310–319, 2024.
- [12] X. Chen, S. Wang, B. Fu, M. Long, and J. Wang, Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning, *Advances in neural information processing systems*, vol. 32, 2019.
- [13] J. Gao, Q. Lao, Q. Kang, P. Liu, L. Zhang, and K. Li, Un-supervised cross-disease domain adaptation by lesion scale matching, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 660–670.
- [14] S. J. Pan and Q. Yang, A survey on transfer learning, *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [15] A. Deng, X. Li, D. Hu, T. Wang, H. Xiong, and C.-Z. Xu, Towards inadequately pre-trained models in transfer learning, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 397–19 408.
- [16] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, Characterizing and avoiding negative transfer, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 293–11 302.
- [17] Y. Ye, Y. Xie, J. Zhang, Z. Chen, Q. Wu, and Y. Xia, Continual self-supervised learning: Towards universal multi-modal medical data representation learning, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11 114–11 124.
- [18] G. Wang, J. Wu, X. Luo, X. Liu, K. Li, and S. Zhang, Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset, *arXiv preprint arXiv:2306.16925*, 2023.
- [19] Y. Jiang, Z. Zhang, S. Qin, Y. Guo, Z. Li, and S. Cui, Apaunet: axis projection attention unet for small target in 3d medical segmentation, in *Proceedings of the asian conference on computer vision*, 2022, pp. 283–298.
- [20] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, Towards a general-purpose foundation model for computational pathology, *Nature medicine*, vol. 30, no. 3, pp. 850–862, 2024.
- [21] Z. Wang, C. Liu, S. Zhang, and Q. Dou, Foundation model for endoscopy video analysis via large-scale self-supervised pre-train, in *International conference on medical image computing and computer-assisted intervention*. Springer, 2023, pp. 101–111.
- [22] Y. Jiang, Z. Zhang, R. Zhang, G. Li, S. Cui, and Z. Li, Yona: You only need one adjacent reference-frame for accurate and fast video polyp detection, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 44–54.
- [23] Y. Jiang, Z. Zhang, Y. Hu, G. Li, X. Wan, S. Wu, S. Cui, S. Huang, and Z. Li, Ecc-polypdet: Enhanced centernet with contrastive learning for automatic polyp detection, *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 8, pp. 4785–4796, 2023.
- [24] J. Jiao, J. Zhou, X. Li, M. Xia, Y. Huang, L. Huang, N. Wang, X. Zhang, S. Zhou, Y. Wang *et al.*, Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis, *Medical image analysis*, vol. 96, p. 103202, 2024.
- [25] Y. Jiang, Y. Hu, Z. Zhang, J. Wei, C.-M. Feng, X. Tang, X. Wan, Y. Liu, S. Cui, and Z. Li, Towards a benchmark for colorectal cancer segmentation in endorectal ultrasound videos: Dataset and model development, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 732–742.
- [26] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, Segment anything in medical images, *Nature communications*, vol. 15, no. 1, p. 654, 2024.
- [27] K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, A generalist vision–language foundation model for diverse biomedical tasks, *Nature medicine*, vol. 30, no. 11, pp. 3129–3141, 2024.
- [28] D. MH Nguyen, H. Nguyen, N. Diep, T. N. Pham, T. Cao, B. Nguyen, P. Swoboda, N. Ho, S. Albarqouni, P. Xie *et al.*, Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching, *Ad-*

- vances in *Neural Information Processing Systems*, vol. 36, pp. 27 922–27 950, 2023.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, Learning transferable visual models from natural language supervision, in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [30] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, Medclip: Contrastive learning from unpaired medical images and text, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [32] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, Uniter: Universal image-text representation learning, in *European conference on computer vision*. Springer, 2020, pp. 104–120.
- [33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, Lora: Low-rank adaptation of large language models. *Iclr*, vol. 1, no. 2, p. 3, 2022.
- [34] D. Lian, D. Zhou, J. Feng, and X. Wang, Scaling & shifting your features: A new baseline for efficient model tuning, *Advances in Neural Information Processing Systems*, vol. 35, pp. 109–123, 2022.
- [35] X. Lei, Y. Ye, Z. Wang, Z. Chen, M. Shu, W. Cai, Y. Zhang, and Y. Xia, Pre-training everywhere: Parameter-efficient fine-tuning for medical image analysis via target parameter pre-training, *arXiv preprint arXiv:2408.15011*, 2024.
- [36] I. Almakky, S. Sanjeev, A. U. R. Hashmi, M. A. Qazi, H. Wang, and M. Yaqub, Medmerge: Merging models for effective transfer learning to medical imaging tasks, *arXiv preprint arXiv:2403.11646*, 2024.
- [37] Y. Yang, L.-K. Huang, and Y. Wei, Concept-wise fine-tuning matters in preventing negative transfer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 753–18 763.
- [38] Y. Yao and G. Doretto, Boosting for transfer learning with multiple sources, in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 1855–1862.
- [39] Y. Yang, B. Nushi, H. Palangi, and B. Mirzasoleiman, Mitigating spurious correlations in multi-modal models during fine-tuning, in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 365–39 379.
- [40] S. Duan, J. Yang, Y. Tan, G. Zhang, Y. Li, and X.-P. Zhang, Transfer risk map: Mitigating pixel-level negative transfer in medical segmentation, in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [41] S. Niu, M. Liu, Y. Liu, J. Wang, and H. Song, Distant domain transfer learning for medical imaging, *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3784–3793, 2021.
- [42] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data*, vol. 6, no. 1, p. 317, 2019.
- [43] K. Yan, X. Wang, L. Lu, and R. M. Summers, Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning, *Journal of medical imaging*, vol. 5, no. 3, pp. 036 501–036 501, 2018.
- [44] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, *arXiv preprint arXiv:2107.02314*, 2021.
- [45] J. N. Kather, N. Halama, and A. Marx, 100,000 histological images of human colorectal cancer and healthy tissue, (*No Title*), 2018.
- [46] C. De Vente, K. A. Vermeer, N. Jaccard, H. Wang, H. Sun, F. Khader, D. Truhn, T. Aimyshev, Y. Zhanibekuly, T.-D. Le *et al.*, Airogs: Artificial intelligence for robust glaucoma screening challenge, *IEEE transactions on medical imaging*, vol. 43, no. 1, pp. 542–557, 2023.
- [47] Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps, in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 387–396.
- [48] Z. Lin, J. Lin, L. Zhu, H. Fu, J. Qin, and L. Wang, A new dataset and a baseline model for breast lesion detection in ultrasound videos, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 614–623.
- [49] N. R. Kurtansky, B. M. D’Alessandro, M. C. Gillis, B. Betz-Stablein, S. E. Cerminara, R. Garcia, M. A. Girundi, E. V. Goessinger, P. Gottfrois, P. Guitera *et al.*, The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection, *Scientific Data*, vol. 11, no. 1, p. 884, 2024.
- [50] D. Kermany, Labeled optical coherence tomography (oct) and chest x-ray images for classification, *Mendeley data*, 2018.
- [51] F. Dernoncourt and J.-Y. Lee, Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts, in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2017, pp. 308–313.
- [52] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images, in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2306–2310.

- [53] E. B. Tsai, S. Simpson, M. P. Lungren, M. Hershman, L. Roshkovan, E. Colak, B. J. Erickson, G. Shih, A. Stein, J. Kalpathy-Cramer *et al.*, The rsna international covid-19 open radiology database (ricord), *Radiology*, vol. 299, no. 1, pp. E204–E213, 2021.
- [54] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang *et al.*, Automated analysis of retinal images for detection of referable diabetic retinopathy, *JAMA ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.
- [55] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in *International conference on medical image computing and computer-assisted intervention*. Springer, 2020, pp. 263–273.
- [56] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, Dataset of breast ultrasound images, *Data in brief*, vol. 28, p. 104863, 2020.
- [57] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.
- [58] M. Subramanian, K. Shanmugavadeivel, O. S. Naren, K. Premkumar, and K. Rankish, Classification of retinal oct images using deep learning, in *2022 international conference on computer communication and informatics (ICCCI)*. IEEE, 2022, pp. 1–7.
- [59] A. Gupta and R. Gupta, Isbi 2019 c-nmc challenge: Classification in cancer cell imaging, *Select Proceedings*, vol. 2, p. 30, 2019.
- [60] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, *arXiv preprint arXiv:1701.06538*, 2017.
- [61] L. Zhang, G. Liu, M. Kong, T. Li, D. Wu, X. Zhou, C. Yang, L. Xia, Z. Yang, and L. Chen, Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model, *Bioinformatics*, vol. 37, no. 11, pp. 1554–1561, 2021.
- [62] X. Lai, J. Zhou, A. Wessely, M. Heppt, A. Maier, C. Berking, J. Vera, and L. Zhang, A disease network-based deep learning approach for characterizing melanoma, *International journal of cancer*, vol. 150, no. 6, pp. 1029–1044, 2022.
- [63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929*, 2020.
- [65] B. Li, X. Xiao, C. Zhang, M. Xiao, and L. Zhang, Dghnn: a deep graph and hypergraph neural network for pan-cancer related gene prediction, *Bioinformatics*, vol. 41, no. 7, p. btaf379, 2025.
- [66] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, Spottune: transfer learning through adaptive fine-tuning, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4805–4814.
- [67] Y. Xia, C. Yang, N. Hu, Z. Yang, X. He, T. Li, and L. Zhang, Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model, *BMC genomics*, vol. 18, no. Suppl 1, p. 950, 2017.
- [68] L. Zhang, W. Bai, N. Yuan, and Z. Du, Comprehensively benchmarking applications for detecting copy number variation, *PLoS computational biology*, vol. 15, no. 5, p. e1007069, 2019.
- [69] W. Wu, L. Song, Y. Yang, J. Wang, H. Liu, and L. Zhang, Exploring the dynamics and interplay of human papillomavirus and cervical tumorigenesis by integrating biological data into a mathematical model, *BMC bioinformatics*, vol. 21, no. Suppl 7, p. 152, 2020.
- [70] L. Zhang, S. Fan, J. Vera, and X. Lai, A network medicine approach for identifying diagnostic and prognostic biomarkers and exploring drug repurposing in human cancer, *Computational and structural biotechnology journal*, vol. 21, pp. 34–45, 2023.
- [71] Y. You, X. Ru, W. Lei, T. Li, M. Xiao, H. Zheng, Y. Chen, and L. Zhang, Developing the novel bioinformatics algorithms to systematically investigate the connections among survival time, key genes and proteins for glioblastoma multiforme, *BMC bioinformatics*, vol. 21, no. Suppl 13, p. 383, 2020.
- [72] Y. You, F. Zhou, Y. Yue, Y. Qiu, X. Wang, Y. Yu, B. Li, R. Li, and L. Zhang, The classical iterative hhl-based hemodynamic simulation quantum linear equation algorithm for abdominal aortic aneurysm, *The European Physical Journal Special Topics*, vol. 234, no. 20, pp. 6253–6263, 2025.
- [73] L. Zhang, W. Song, T. Zhu, Y. Liu, W. Chen, and Y. Cao, Convnext-mhc: improving mhc-peptide affinity prediction by structure-derived degenerate coding and the convnext model, *Briefings in Bioinformatics*, vol. 25, no. 3, p. bbae133, 2024.
- [74] B. Li, M. Xiao, R. Zeng, and L. Zhang, Developing a multiomics data-based mathematical model to predict colorectal cancer recurrence and metastasis, *BMC Medical Informatics and Decision Making*, vol. 25, no. Suppl 2, p. 188, 2025.
- [75] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, Medvit: a robust vision transformer for generalized medical image classification, *Computers in biology and medicine*, vol. 157, p. 106791, 2023.
- [76] G. J. Chowdary and Z. Yin, Med-former: A transformer based

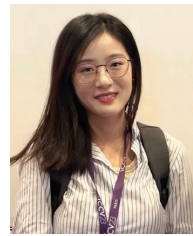
architecture for medical image classification, in *International conference on medical image computing and computer-assisted intervention*. Springer, 2024, pp. 448–457.

- [77] A. S. Remigio, Incarmag: A convolutional neural network with multi-level autoregressive moving average graph convolutional processing framework for medical image classification, *Neurocomputing*, vol. 617, p. 129038, 2025.
- [78] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [79] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [80] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, *arXiv preprint arXiv:2102.04306*, 2021.
- [81] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in *European conference on computer vision*. Springer, 2022, pp. 205–218.
- [82] W. Fedus, B. Zoph, and N. Shazeer, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
- [83] J. Seo, J. Kim, and H. Shin, Mofe: Mixture of frozen experts architecture, in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 2025, pp. 340–348.

Author biography



Yuncheng Jiang received his PhD in Computer and Information Engineering from The Chinese University of Hong Kong, Shenzhen. He previously served as a Visiting Researcher at A*STAR, Singapore, and is currently a Postdoctoral Fellow at West China Hospital, Sichuan University. He has served as a reviewer for multiple prestigious conferences and journals including MICCAI, TMI, CVPR, and ICIC. His research primarily focuses on interdisciplinary medical artificial intelligence, with core directions in deep learning-based lesion segmentation and detection, and multimodal imaging-assisted cancer diagnosis. He has applied for 1 national invention patent. During his PhD studies, he published 6 first-author papers in top-tier journals and conferences including npj Digital Medicine, IEEE JBHI, and MICCAI, with a total of 15 high-quality papers published. His current research interests include medical artificial intelligence, deep learning-based medical image analysis, and multimodal imaging for cancer diagnosis.



Chun-Mei Feng received her Ph.D. in Computer Science from Harbin Institute of Technology (Shenzhen) in 2022. She completed an internship at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in Abu Dhabi in 2020, followed by a visiting scholar position at ETH Zurich in 2021. She previously served as a Research Fellow at A*STAR Singapore and a Visiting Researcher at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia. Currently, she is an Assistant Professor at the School of Computer Science, University College Dublin (UCD), where she is also a UCD Ad Astra Fellow, and concurrently holds a Visiting Researcher position in the Department of Biostatistics and Health Informatics at King's College London. Her research has been published in over 60 papers in prestigious journals and conferences, including Nature Communications, npj Digital Medicine, TPAMI, IJCV, TMI, TNNLS, TIP, CVPR, ICCV, ECCV, ICLR, MICCAI, and AAAI. Her current research interests focus on medical artificial intelligence, multimodal learning, and federated learning.



Jiankun Hu is Chief Physician of Gastrointestinal Surgery, Director of the Gastric Cancer Research Laboratory, Deputy Director of the Gastrointestinal Surgery Center, and Secretary of the Party General Branch of Surgery at West China Hospital, Sichuan University. As a Fellow of the Royal College of Surgeons, he has received honors including the Program for New Century Excellent Talents in University and leads multiple national and provincial research projects. He is an expert reviewer for the National Health Commission and National Natural Science Foundation. A leading gastrointestinal oncology surgeon, he proposed the 'forward modular laparoscopic gastric cancer lymph node dissection' technique and led the 2017 Chinese Expert Consensus on laparoscopic gastric surgery quality control. His research on gastric cancer stem cells and the CLDN18-ARHGAP26 fusion gene has been published in Cell Research and Nature Communications. He has published over 160 papers (100+ SCI-indexed) and supervised 35 graduate students. His current research focuses on gastric cancer pathogenesis, minimally invasive treatment, and cancer stem cells.



Lusheng Wang is a faculty member in the Department of Computer Science, City University of Hong Kong. His research interests include algorithms, bioinformatics, computational biology and artificial intelligence. He has authored hundreds of papers published in renowned journals such as TCBB, Bioinformatics, JCSS and JACM. He serves as an associate editor for IEEE/ACM TCBB, BMC Bioinformatics, the International Journal of Knowledge Discovery in Bioinformatics and JCSS, and is also on the editorial boards of the Journal of Bioinformatics and Computational Biology and the Journal of Global Optimization. His current

research focuses on Algorithms in Bioinformatics.



Le Zhang received his PhD in Computer Science from Louisiana Tech University in 2005. He served as a Postdoctoral Fellow at Harvard Medical School and held faculty positions at Michigan Technological University, University of Rochester, and Southwest University. He is currently Professor and Associate Dean

at College of Computer Science, Sichuan University. He has been honored with the Chongqing Hundred Overseas High-Level Talent Program and Sichuan Provincial Academic and Technical Leader. Dr. Zhang has led major research projects funded by NIH and NSFC, and received awards including the Wu Wenjun Artificial Intelligence Second Prize and Sichuan Science and Technology Progress Second Prize. With over 150 SCI-indexed publications, his research focuses on AI-based analytical tools for medical and biological research, including medical imaging and natural language analysis.

Just Accepted