Optimizing Your Foundation Model for Medical Images: A Comprehensive Analysis of Fine-Tuning Strategies

Anonymous ACCV 2024 Submission

Paper ID #1106

Abstract. Recent studies have shown the potential of fine-tuning foundation models on downstream natural images, resulting in impressive performance. However, our investigations reveal that fine-tuning foundation models on medical images leads to highly variable performance, where no single fine-tuning strategy emerges as universally superior across all tasks. The observed variation in performance can be attributed to disparities between natural and medical images, as well as distinctions in data dimensionalities. This implies that researchers must rely on a time-consuming trial-and-error approach to identify the optimal fine-tuning method from a plethora of recently developed techniques. To address this issue, we propose a simple yet effective solution called Modality-Understanding Tuning (μ-Tuning) that automatically adjusts the tuning mechanism when data modality is different. Through a comprehensive analysis involving seven state-of-the-art fine-tuning strategies and twelve diverse medical imaging datasets, we investigate their respective strengths and weaknesses. Our results demonstrate that μ-Tuning outperforms alternative fine-tuning strategies and can serve as an effective one-stop shop for adapting a pre-trained foundation model to downstream medical imaging applications. Source code is publicly available at here.

Keywords: Transfer Learning · Medical Applications · Visual Prompt Tuning · Minimal Weight Tuning.

1 Introduction

Medical imaging applications heavily depend on meticulously curated data and annotations to facilitate effective analysis and diagnosis. Nevertheless, the scarcity of labeled data and annotations presents a significant hurdle in the creation of robust and precise models for medical image analysis. This challenge is magnified due to the labor-intensive nature of obtaining large, well-annotated datasets, a process that demands expert knowledge and manual involvement, especially for intricate tasks that involve large 3D imaging datasets.

Transferring knowledge from pre-trained or foundation models has emerged as the predominant strategy to address the issue of limited labeling in medical analysis applications, especially with the introduction of foundation models like SimCLR [14], MoCo [26], MAE [25], CLIP [48], and SAM [34]. However,

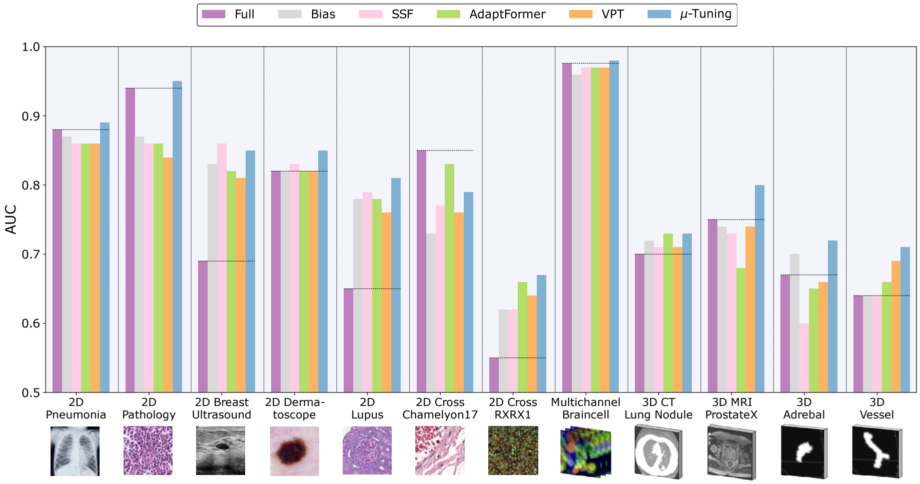

Fig. 1: Performance comparison between μ-Tuning and state-of-the-art methods in few-shot scenario (100 samples per class) across a diverse range of 2D and 3D tasks. μ-Tuning significantly outperforms different baselines including FULL tuning, BIAS [8,66], SSF [37], AdaptFormer [13], and VPT [31]. While existing fine-tuning techniques lack consistency across all tasks. Best viewed in color.

the conventional approach of transfer learning, which entails fine-tuning the entire pre-trained model or merely the final layer, can encounter the challenge of *catastrophic forgetting* [23,43,49]. This phenomenon refers to the risk that retraining or fine-tuning a pre-trained model for a novel task might lead to the degradation of previously acquired knowledge [43,58].

To overcome this challenge, various techniques have been developed over time to harness the robust representations of foundation models effectively. Notably, the techniques of prompt tuning and minimal parameter tuning have gained considerable attention. *Prompt-tuning* emerges as a promising solution to address the challenges of few-shot learning for medical image analysis [31,51,58]. By fine-tuning a small set of *additional* input tokens, this technique reshapes the target domain input to optimally align with the frozen pre-trained models. Conversely, *minimal weight-tuning* strategies, such as Bias [8,66], SSF [37], Adapter [13,29,30], propose adjusting only a limited subset of parameters within the model, while keeping the rest unaltered.

Given the substantial potential demonstrated by these tuning methodologies on natural images, a natural question arises: *How effective are these approaches in fine-tuning medical imaging tasks across diverse modalities?* This question underscores the need for a systematic investigation and comprehensive evaluation to shed light on the suitability of these techniques for medical image analysis. According to our findings, neither prompt tuning nor weight-tuning excels across all datasets, see Fig. 1.

In this paper, we conduct a comprehensive investigation of the existing state-of-the-art fine-tuning methods for medical imaging tasks. Motivated by the insights from this investigation, we introduce a novel Modality-Understanding Tuning (μ-Tuning) capable of adjusting the tuning complexity based on the imaging modality. μ-Tuning enhances the adaptability and performance of models in few-shot scenarios, especially when going beyond 2D imaging tasks. In summary, the contributions of this study are as follows:

- Developing μ-Tuning, a novel, unified tuning framework capable of working with both 2D and 3D medical imaging modalities, exhibiting superior stability and performance compared to existing fine-tuning methods.
- Conducting a comprehensive performance analysis of visual prompt and weight tuning techniques in few-shot medical image analysis scenarios.

The rest of the paper is organized as follows: Section 2 discusses related works; Section 3 provides background knowledge on vision transformers, visual prompting, and minimal-weight tuning to facilitate understanding of this paper; Section 4 provides the motivation and technical description of μ-Tuning; Section 5 presents extensive experimental results and discussion. Finally, Section 7 concludes the paper with future research directions. In addition, further results and discussions are introduced in the *Supplementary Material*.

2 Related Works

Visual Prompting in Medical Imaging. VPT [31] has recently emerged as a promising technique for fine-tuning pre-trained models for downstream tasks. Notably, VPT has been applied to diverse tasks, including visual prompt-based adversarial attacks [5, 11], domain generalization [68, 71], long-tailed prompt tuning for image classification [18], video applications, such as tracking [55, 61], generative visual prompts [6, 60, 62], visual question answering (VQA) [28, 62], continual learning [55, 57, 58], and point cloud analysis [59, 67]. Recent works have investigated VPT for two-dimensional medical image analysis. Noteworthy efforts are visual prompt-tuning for medical imaging [24], visual prompt-based tuning for head and neck cancer segmentation [50], and visual prompt-based federated learning for MRI reconstruction [22]. Additionally, [16, 47] focus on adapting large vision-language foundation models to analyze medical images and dealing with long-tailed medical image distributions across datasets [21]. However, current research on visual prompt tuning for medical applications primarily focuses on individual applications, predominantly utilizing 2D imaging modalities. This leaves a significant gap as 3D images (such as CT and MRI) and multi-dimensional images (e.g., multiplexed fluorescence images) are crucial for disease diagnosis, treatment planning, and drug discovery.

Minimal Weight Tuning in Medical Imaging. Minimal Weight Tuning strategies have gained significant traction in the scientific community due to their ability to selectively modify a subset of parameters in a pre-trained model, leaving the majority of the model intact. These strategies can be classified based on their signal transformation techniques, which span both linear and non-linear modalities. Within the linear tuning domain, the bias tuning methodology [8, 66] is dedicated to the fine-tuning of bias terms exclusively in pre-trained models. SSF [37] incorporates scale and shift transformations to the outputs of pivotal operations in the Vision Transformer, addressing certain constraints inherent to bias tuning.

Regarding the non-linear tuning, the Adapter [29,30] stands out as a prominent technique. This approach rectifies a significant drawback of Vision Transformers, where the output from self-attention mechanisms tends to diminish exponentially as network depth increases [19,56]. The Adapter technique integrates a bottleneck-inspired module as a residual connection within each layer, thereby augmenting the feature expressiveness. Several iterations of the Adapter methodology have been proposed, such as AdaptFormer [13], ViT-Adapter [15] and [12], each tailoring the foundational principle to diverse visual tasks and architectures.

Despite the promising results exhibited by Minimal Weight Tuning techniques in comparison to Visual Prompt Tuning, their usage in the biomedical domain remains relatively under-explored [24]. In addition, the efficacy of these tuning strategies on 3D and higher-dimensional modalities remains an open research question. Recognizing this research gap, our paper comprehensively analyzes the efficacy of various fine-tuning strategies on a wide range of data modalities. Our experiments cover diverse imaging tasks in the brain, breast, lung, and prostate. We also introduce a unified framework, called μ-Tuning, capable of working with a wide range of data modalities, spanning both 2D images to higher-dimensional modalities such as CT, MRI, X-ray, ultrasound, and microscopy.

3 Background

Vision Transformer (ViT) [20] is a deep learning model architecture that has been primarily developed for image classification tasks. Formally, for a ViT with N layers, the input image is divided into m number of fixed-sized image patches I and then embedded into a d-dimensional latent space with positional encoding. The mathematical formulation of ViT is as follows:

$$X^j = \text{Embed}(I^j), \qquad X^j \in \mathbb{R}^d, j = 1, 2, ..., m \tag{1}$$

$$[c_i, X_i] = L_i([c_{i-1}, X_{i-1}]),\ X_i = \{x_i \in \mathbb{R}^d | i = 1, ..., N\} \tag{2}$$

$$y = \text{Head}(c_N) \tag{3}$$

where, *Embed* layer transforms image patches into tokens. The $[\cdot, \cdot]$ represents the concatenation of tokens and $c_i \in \mathbb{R}^d$ denotes the output embedding token corresponding to the class token *[CLS]* at layer L_i. Each layer L_i consists of Multi-headed Self-Attention (MSA), Feed-Forward Networks (FFN), LayerNorm [4], and residual connections [27]. The mathematical formulations for Self-Attention and Multi-Head Self-Attention are given below:

$$\text{Attn}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V \tag{4}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{Attn}_1, \text{Attn}_2, ..., \text{Attn}_{12}) \tag{5}$$

The Multi-headed Self-Attention consists of multiple Self-Attention modules. In each Self-Attention module, the Q, K, and V denote the query, key, and value matrices, respectively. The attention mechanism computes the relevance scores between the queries and keys and uses them to weigh the corresponding values. The division by $\sqrt{d}$ is a scaling factor that helps stabilize the training process [54]. Finally, the output of the attention mechanism is computed as the weighted sum of the values, representing the self-attended token embeddings for that layer.

Fine-Tuning Methodologies In this section, we provide technical descriptions of popular strategies for fine-tuning foundation models, including VPT [31], SSF [37], and Adapter [13, 29, 30, 45, 46]. We highlight the tunable parameters with $\widetilde{\text{tilde}}$.

- *Full Tuning.* This strategy is one of the most popular transfer learning methods where all the model is updated during the fine-tuning process to adapt the pre-trained model to a new task. However, it is prone to overfit on the currently available data and suffers from performance deterioration on the previously trained data due to catastrophic forgetting, especially on few-shot learning scenarios [7, 43, 57, 58].
- *Classification Head Tuning* This strategy [13, 31, 37] is the process of freezing the entire model and only fine-tuning the classification head. However, this technique often suffers from data distribution and modality shifting as the model's weights are not updated. Formally, the mathematical formulation of the classification head tuning is as follows:

$$x_{out} = (x_{in} \odot \widetilde{A^T} + \widetilde{b}) \tag{6}$$

  where (x_{in}, x_{out}) are respectively input and output of a network's classification layer, and $(\widetilde{A^T}, \widetilde{b})$ are tunable parameters, $\odot$ is the dot product.
- *Visual Prompt-Tuning* (VPT) [31] is a fine-tuning technique that enhances the capabilities of pre-trained ViT models by incorporating tunable embedding vectors, referred to as prompt tokens. These tokens are introduced to the input space of each layer, enabling the model to adapt to the downstream tasks. The VPT mechanism is formally presented as:

$$[c_i, x_i] = L_i[c_{i-1}, \widetilde{P}_{i-1}, x_{i-1}] \tag{7}$$

$$y = \text{Head}(c_N) \tag{8}$$

  VPT introduces a collection of tunable prompt tokens, into the input space of each layer L_i. These tokens are represented by the matrix $P_i \in \mathbb{R}^{d \times p}$, where d is the dimension of the embeddings, and p is the number of prompt tokens. The tunable prompts are stacked alongside the original tokens x_{i-1}, creating an augmented input, i.e., $[c_{i-1}, P_{i-1}, x_{i-1}]$. During fine-tuning, only the prompt tokens are updated and the rest of the model is frozen, ensuring that the pre-trained knowledge is retained.
- *Bias Tuning* [8, 66] involves fine-tuning only the bias terms of the pre-trained model, denoted as $\widetilde{b}$ in Eq. 9. In this method, while the pre-trained weights of the model are frozen and remain unchanged, the bias terms are updated during the fine-tuning process. Bias tuning is computationally efficient and requires fewer resources compared to full weight-tuning methods since it involves updating only a small portion of the model's parameters. However, it may not

be as effective in adapting to new datasets with significant distribution shifts, especially in scenarios where the data characteristics differ significantly from the original pre-training dataset.

$$\text{Bias Tuning: } x_{out} = x_{in} \odot A^T + \widetilde{b} \tag{9}$$

$$\text{Scale and Shift Fusion: } x_{out} = (x_{in} \odot A^T + b)\widetilde{\alpha} + \widetilde{\beta} \tag{10}$$

Note that in a ViT block, there are multiple operations such as QKV projections, and linear mappings. Each of these operations associate with a bias term. Subsequently, there are multiple fine-tuning components within a single ViT block. With this in mind, x_{in} and x_{out} are not simply the input and output of each ViT layer, but input and output of each operation in a ViT layer.

- *Scale and Shift Fusion* (SSF) [37] aims to address the limitations of bias tuning by introducing scale and shift transformations to the output of key operations within the ViT. Specifically, SSF applies the scale and shift transformations to the output of important operations such as the multi-head self-attention (MSA) module, the MLP, and the layer normalization (LN) step. The scale and shift transformations are controlled by tunable parameters $\widetilde{\alpha}$ and $\widetilde{\beta} \in \mathbb{R}^d$, as shown in Eq. 10, where A and b are the weight and bias terms, respectively. The $\odot$ denotes the dot product. By introducing these tunable parameters, SSF allows for some adaptation to new datasets and variations in data distributions. However, like bias tuning, SSF employs linear transformations, which may not be sufficient to handle complex distribution shifts, especially in the case of 3D data where the data properties may differ significantly from those of 2D data.
- *Adapter* [13, 29, 30] takes a more comprehensive approach to minimal weight-tuning by introducing an additional MLP module with a non-linear transformation, such as ReLU, along with a residual connection to the original Feed-Forward Networks (FFN) inside each layer of the ViT, as represented in Eq. 11 and 12. By adding the non-linearity, the adapter approach demonstrates improved adaptability to new datasets, especially when facing distribution shifts. The adapter module focuses on enhancing the MLP part of the ViT, which has been shown to be particularly crucial in addressing the rank collapse problem in ViT [19, 56]. The rank collapse refers to the phenomenon where the output of self-attention and/or output of ViT collapses to a rank-1 matrix, limiting the model's expressive capacity. By leveraging adapters, the model can better capture long-range dependencies and complex patterns in the data, facilitating superior generalization to new domains.

$$\hat{x}_{out} = \phi\big(\text{LN}(x_{in}) \odot \widetilde{\mathcal{W}}_{down}\big) \odot \widetilde{\mathcal{W}}_{up} \tag{11}$$

$$x_{out} = \text{MLP}\big(\text{LN}(x_{in})\big) + \widetilde{s}_{ada}\, \hat{x}_{out} \tag{12}$$

where $\widetilde{s}_{ada} \in \mathbb{R}$ maps the output of the adapter block to the scale of FFN's output.

Comparing these four tuning methods, we can observe that bias tuning and SSF offer simplicity and efficiency by updating only specific parameters. On the other hand, the Adapter approach introduces non-linearity through the additional MLP module, which enables more effective adaptation to diverse and

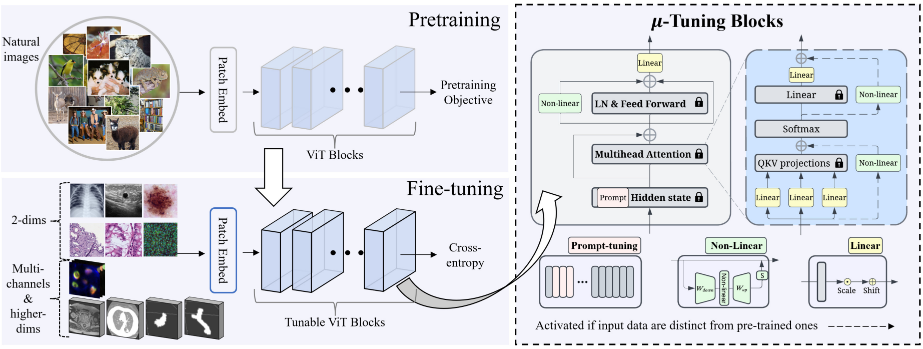

Fig. 2: Overview of μ-Tuning. In particular, to better address the linear distribution shift which has similar two-dimensional data, we insert Linear module into ViT's block backbone to consequently align the output of the preceding operation with the scale and range of the input space of the subsequent operation. In addition, since inadequate tuning might greatly hinder model performance in high-dimensional tasks that are distinct from pre-trained data, μ-Tuning introduces non-linearity into ViT's block by utilizing a simple non-linear bottleneck-like MLP, referred to as Non-Linear . We also introduce a set of tunable embeddings into the input space of each layer within ViTs. In contrast to Linear and Non-Linear tuning modifying the backbone, the Prompt-tuning updates a set of learnable input tokens which may reshape the new input to optimally align with the pre-trained models. In the case of high-dimension tasks, re-weighting the *Patch Embedding* module is needed. Best viewed in color.

complex data distributions. Similarly, VPT introduces tunable embeddings and non-linear transformations. However, each tuning method has its own limitations. We hypothesize that combining these methods can improve the robustness and performance of ViT models across various computer vision tasks, especially when dealing with challenging high-dimensional data or datasets with unique characteristics.

4 Modality-Understanding Tuning

In this section, we formally present a novel strategy, named Modality-Understanding Tuning (MU or μ-Tuning) capable of adjusting the tuning complexity based on the input modality. This adaptive tuning strategy is key to overcoming the limitations of both prompt tuning and minimal weight tuning strategies.

Linear and Non-Linear Tuning Inspired by the Adapter and Bias Tuning, we define general linear and non-linear tuning blocks to facilitate the discussion of μ-Tuning architecture and comparison between the two approaches. Our definitions of *Linear* and *Non-Linear* blocks are general in the sense that they can be applied to various operations in the network. In contrast, previous methods like the Adapter [13] are limited to applying them to the residual connection in ViT blocks or unspecific applied throughout ViT blocks as Bias [8, 66]. The mathematical representations of Linear, Non-Linear, and Prompt-tuning mechanisms are defined as follows:

$$\begin{aligned} \text{Linear-Tuning:}\quad & x_{out} = (x_{in} \odot A^T + b)\widetilde{\alpha} + \widetilde{\beta} && (13)\\ \text{Non-Linear-Tuning:}\quad & x_{out} = \phi\big(x_{in} \odot \widetilde{\mathcal{W}}_{down}\big) \odot \widetilde{\mathcal{W}}_{up} && (14)\\ \text{Prompt-Tuning:}\quad & x_{out} = Layer[\widetilde{P}_{in}, x_{in}] && (15) \end{aligned}$$

More details of our architecture are provided in the subsequent section. Since our strategy embeds linear and non-linear tuning blocks in a more comprehensive fashion across various operations within ViT, our method can adapt to more complex changes in data distribution.

4.1 μ-Tuning Architecture

This method aims to address several limitations of existing tuning methods. First, previous methods are quite rigid and unable to cope with the changes in the data distribution. This is especially detrimental when it comes to changes caused by discrepancies in data modalities between the pre-trained and adapted ones. Second, the tuning parameters in previous methods are not embedded extensively throughout different operations in networks.

In contrast, our method adaptively modifies various operations, leading to better adaptation capability. We hypothesize that tasks involving two-dimensional medical imaging might share a similar data distribution with the datasets used for pre-training foundation models. Conversely, tasks involving higher-dimensional data are likely to deviate significantly from the pre-training distribution. μ-Tuning addresses the limitations of previous methods and aims to adapt to these data distribution shifts by integrating both *Linear* and *Non-Linear* tuning within the Vision Transformers (ViTs) backbone. Figure 2 provides an overview of the μ-Tuning architecture. The mathematical formulation of μ-Tuning are as below:

$$\begin{aligned} x_{out} &= Layer[\widetilde{P}_{in}, x_{in}] && (16)\\ [Q, K, V] &= [(x\widetilde{\alpha}_{q/k/v} + \widetilde{\beta}_{q/k/v})A_{q/k/v} + b_{q/k/v}] && (17)\\ x_{out} &= x_{in}W + (\phi(x \odot \widetilde{\mathcal{W}}_{down}) \odot \widetilde{\mathcal{W}}_{up})\mathbb{I}_{dim} && (18)\\ \text{Attn}_i(Q, K, V) &= \text{Softmax}(QK^T/\sqrt{d})V && (19)\\ \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{Attn}_1, \text{Attn}_2, ..., \text{Attn}_{12}) && (20)\\ \text{MSA}_{out} &= \text{MLP}(\text{MultiHead}(Q, K, V))\\ &\quad + (\phi(\text{MultiHead}(Q, K, V) \odot \widetilde{\mathcal{W}}_{down}) \odot \widetilde{\mathcal{W}}_{up})\mathbb{I}_{dim} && (21) \end{aligned}$$

Here, $\widetilde{\alpha}_{q/k/v}$ and $\widetilde{\beta}_{q/k/v}$ are tunable parameters for Q, K, and V, respectively. $\mathbb{I}_{dim}$ is an indicator function that turns on when the dimension of the input data is different from that of the pre-trained data.

The proposed architecture has several novelties. First, to adapt to two-dimensional tasks that potentially have similar data distribution to the pre-trained one, the Linear Tuning transformation is applied to the input of Q, K, and V projection of MSA as shown in Eq. 17. The goal is to consequently modify the attention operation. More precisely, let $x \in \mathbb{R}^{m \times d}$ denote the sequence input of fixed-sized image patches, where m is the number of image patches and d is the embedding dimension. Let $W_q, W_k, W_v \in \mathbb{R}^{d \times h}$ denote the the projection matrices, where h is the hidden dimension. These matrices map x to the queries Q, keys K, and values V, where $Q, K, V \in \mathbb{R}^{m \times h}$. In the fine-tuning process where the projection matrix W is fixed, changing x would effectively change the dot product $x^T W$. As a result, it allows the MSA to pay attention to different regions with the same input x in every epoch while still having the ability to preserve the pre-trained knowledge of W_q, W_k, W_v if needed by setting the scale $\widetilde{\alpha} \approx 1$ and shift $\widetilde{\beta} \approx 0$.

Second, we introduce more non-linearity into the input of MSA operations. We hypothesize that inadequate tuning might impede model performance in high-dimensional tasks. Thus, the additional Non-Linear module is particularly valuable for adapting pre-trained models on three-dimensional (3D) datasets, where data distributions significantly differ from 2D natural images used for training foundation models. This change is reflected in Eq. 18, Eq. 21, and illustrated in Fig. 2. Specifically, the non-linearity is added using two MLPs to compress and upsample the features in the residual connections of the MLP and Multiheaded Self-Attention (MSA) operations in the pre-trained ViT. Here, $\mathcal{W}_{down} \in \mathbb{R}^{d \times r}$ and $\mathcal{W}_{up} \in \mathbb{R}^{r \times d}$ are projection matrices, and $\phi(\cdot)$ denotes a non-linear activation function, such as ReLU. The Eq. 19 and 20 are the same Softmax and Concatinate operations in original ViTs.

Finally, μ-Tuning addresses the complexity associated with both 2D and high-dimensional tasks in a unified manner. Our experiments demonstrate that μ-Tuning's achieves superior performance across diverse 2D and 3D imaging datasets in few-shot settings. Furthermore, it enhances training stability, efficiency, and applicability across various medical imaging modalities. In this study, we employ the VPT [31], SSF [37], and Adapter [13,29,30,45,46] for prompt-tuning, linear, and non-linear tuning modules, respectively. The reason we chose these methods is that they could be easily replaced by their superior variants for further research and specific use cases.

Interestingly, GLoRa [10] and NOAH [69] also propose the integration of VPT, LORA, and Adapter, conducting neural search on hidden dimensions, rank, and prompt length. However, our work fundamentally differs from GLoRa and NOAH in two aspects. First, we propose applying scaling and shifting modulation at the input of QKV projection, re-weighting the projection without direct fine-tuning. Second, our introduction of the Adapter block as a residual connection to QKV projection before the Softmax operation is specifically designed to introduce non-linear tuning signals for adaptation beyond two-dimensional data. Conversely, in NOAH, the residual connection to QKV projection involves LORA, which serves to prevent self-attention output collapse to a rank-1 matrix [19, 56]. Furthermore, this study aims to address the intricate challenge of adapting pre-trained models from natural images to the medical domain, particularly in contexts

with limited data availability. These reasons separate our work from the other transfer learning methodology and medical application papers.

To utilize the 2D pre-trained models for 3D medical images, we employ an effective *weight average inflation strategy* [9, 63, 70]. Please refer to the *Supplementary Material* for details.

5 Experimental Settings

Pre-Trained Backbones We experiment with pre-trained ViT [20] backbone from the vision branch of CLIP [48]. Since we use the backbone from the CLIP model, we follow the architecture configurations, e.g., patch size, number of layers, number of attention heads, etc. We choose to utilize the CLIP [48] vision pre-trained weights because its connection with the corresponding languages' pre-trained weights and opening up the possibility of incorporating μ-Tuning into other vision-language research and applications such as LLaVa [39, 40] and LLaVa-Med [36]. We also add a classification head to the backbone of all the experiments.

Datasets We experiment with a wide range of imaging types and clinical-purpose tasks. Note that some datasets involve high-dimensional data samples and can not directly leverage the 2D pre-trained foundation models without significant modifications to fine-tuning methodologies. The two-dimensional tasks consist of seven datasets, including Pneumonia pediatric chest X-Ray [33], colorectal cancer histology patches [32], cancer breast ultrasound [1], dermatoscopic images [17, 53], Lupus (private), cross-domain Camelyon17 Patch [35], and cross-domain RXRX1 [35, 52]. The high-dimensional tasks include multi-channel braincell [41, 42], CT lung nodule [3], MRI cancer prostate [2, 38], CT adrenalmnist [64], and reconstructed MRA vesselmnist [64, 65]. Refer to the *Supplementary Material* for details.

Evaluation Metrics Given that medical datasets are often highly class-imbalanced, we focus our evaluation on the Area Under the Curve (AUC) rather than accuracy (ACC). For completeness, we also include the ACC results in the *Supplementary Material*. The AUC and ACC reported for each dataset represent the mean outcomes from three separate runs, each initialized with a distinct random seed.

6 Results and Discussions

In this section, we provide a comprehensive analysis of the performance of all fine-tuning methods discussed in Sections 3. To provide statical insight, we also present the McNemar's test in section 6.2. In addition, we also introduce the ablation between linear versus non-linear tuning in μ-Tuning in section 6.3. Furthermore, we present other analyses, such as *cross-domain performance*, *stability analysis*, *parameter efficiency*, and *accuracy performance* in *Supplementary Material*.

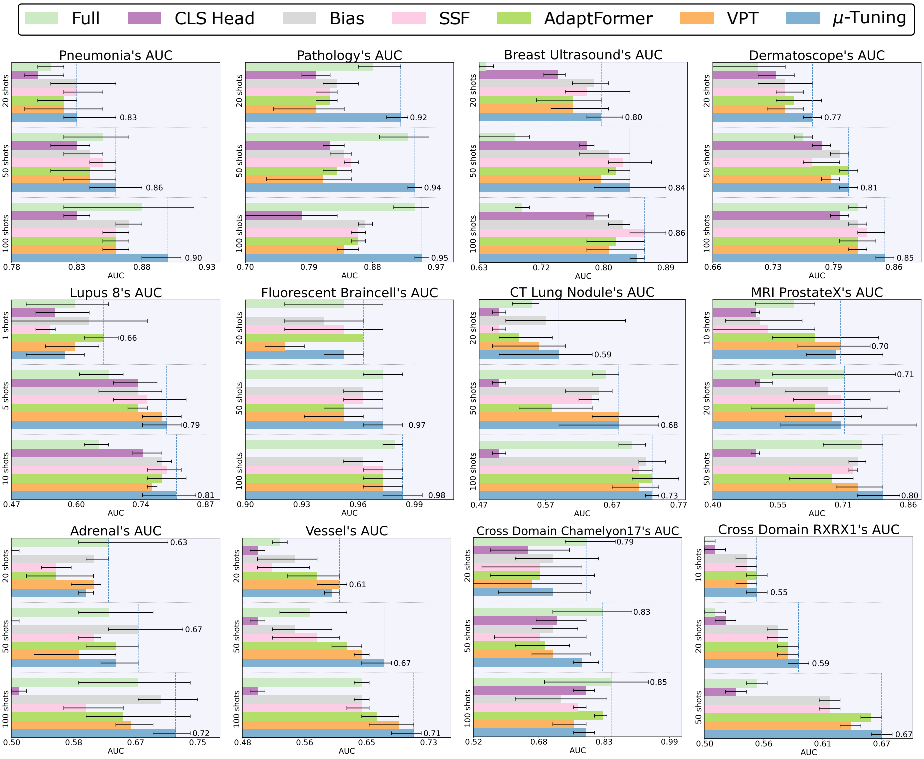

Fig. 3: Comparisons of finetuning AUC results of μ-Tuning and state-of-the-art transfer learning methods on twelve distinct medical imaging tasks and seven two- and high-dimensional modalities. Note that, to ease the visualization, we only show the highest averaged scores of each setting.

6.1 Performance Comparison

Fig. 3 presents a comprehensive comparison of all fine-tuning methods. Generally, *BIAS* [8, 66], *SSF* [37], *AdaptFormer* [13], and *VPT* [31] show promising yet inconsistent performances across tasks. The high variability in the results of these methods could be due to various factors, including input token length in VPT [31], the hidden dimension in the Adapter block [13, 29, 30], and the manual setting of which ViT layers and operations remain unchanged while applying *BIAS* or *SSF* to each task. The sensitivity in parameter selection that leads to performance inconsistency has been acknowledged in the original papers [8, 13, 37, 66]. Interestingly, *VPT* is inferior to simpler methods, such as *BIAS* and *SSF*, in six out of twelve datasets. This observation highlights the significant challenge of adapting 2D foundation models to medical tasks.

In contrast, μ-Tuning consistently achieves superior performance in eleven out of twelve datasets under different few-shot settings. For 2D tasks, μ-Tuning outperforms *FULL* tuning while utilizing substantially fewer parameters, see *Supplementary Material* for details. The *catastrophic forgetting* problem associated with fine-tuning the entire network with limited data [43, 58] renders *FULL*

Table 1: The p-value from McNemar' test [44] and the ratio B/C between our μ-Tuning and other state-of-the-art methods. The p-value that is smaller than the significance threshold, e.g., $\alpha = 0.05$, suggests the two model's performances on a specific task are not equal. Regarding the ratio B/C ratio from the contingency table, values that are greater than 1.0 indicate that the number of samples that our μ-Tuning correctly predicts while the other is wrong is higher than the number of samples that our μ-Tuning wrongly predicts while the other is right. The comparisons have a p-value smaller than 0.05 or the ratio B/C higher than 1.0 is shaded in cyan color. These results indicate our μ-Tuning achieves statistically significant improvement over other baselines.

Metrics	Method	Chest XRay	Colon Pathology	Breast Ultrasound	Derma-toscope	Lupus 8	Came-lyon17	RXRX1	CT Lung	MRI ProstateX	Multi-channel Braincell	CT Adrenal	MRA Vessel
p-value	Bias	0.005	0.0	0.009	0.037	0.02	0.0	0.0	0.003	0.0	0.31	0.003	0.0
	SSF	0.0	0.0	0.3	0.0	0.62	0.229	0.0	0.07	0.055	0.1	0.01	0.001
	AdaptFormer	0.003	0.0	0.001	0.0	0.14	0.945	0.04	0.025	0.066	0.02	0.33	0.0
	VPT	0.28	0.0	0.832	0.001	0.18	0.07	0.0	0.02	0.5	0.01	0.01	0.14
B/C Ratio	Bias	2.14	2.29	3.33	1.25	0.23	0.54	2.37	1.66	1.20	3.12	2.16	1.76
	SSF	3.94	6.89	1.67	3.72	1.43	0.92	2.33	1.61	1.55	1.54	2.5	2.12
	AdaptFormer	2.26	1.38	4.80	1.89	0.42	1.00	1.1	1.55	2.47	1.57	1.73	2.54
	VPT	1.47	1.43	1.20	0.69	0.29	1.37	1.69	1.42	1.18	1.70	1.95	1.4

tuning less effective in most 2D tasks under the lowest-shot settings. Conversely, μ-Tuning leverages selective tuning to ensure that important pre-trained features are preserved while adapting to new tasks. For 3D tasks, it is interesting that *FULL* outperforms *LINEAR* and *SSF*, while being competitive with *BIAS*, *ADAPTFORMER*, and *VPT*. One reason could be that adapting 2D pretrained models to 3D data requires a more extensive adaptation of the model due to the significant difference in the feature space. As a result, tuning more parameters like in *FULL*, including those within MSA, MLP, and LN, leads to higher performance. This observation corroborates the importance of adding non-linearity to the case of 3D adaptation, as done automatically in μ-Tuning.

Visual Prompt Tuning (VPT) [31] technique involves modifying the input for new tasks by adding an extra sequence of tokens to align with the pre-trained model's input space. Although VPT achieves competitive performance on 2D tasks by leveraging the pre-trained representations of CLIP [48] and visual prompt tokens of MVLPT [51], μ-Tuning surpasses the performance of VPT on both 2D and 3D tasks with significant margins. Furthermore, our results indicate that while the mechanism of adapting input prompts in VPT is effective, μ-Tuning offers enhanced stability and consistency. The superior stability exhibited by μ-Tuning makes it especially suitable medical applications where consistent outcomes are crucial, as discussed further in *Supplementary Material*. The instability of VPT performances could be attributed to the challenge of determining optimal input token lengths for each downstream task.

Minimal Weight Tuning includes methods such as *BIAS* [8,66], *SSF* [37], and *ADAPTFORMER* [13], which tune only a subset of pre-trained model parameters while keeping the rest frozen. In 2D tasks, these methods leverage the strong pre-trained representations of CLIP [48] and achieve competitive results by fine-tuning a small subset of parameters. However, μ-Tuning outperforms them in all tasks, where a method is considered superior in a task if it performs the best in the majority of few-shot settings. In 3D tasks, *BIAS* and *SSF* apply linear tuning to key operations like MLP, MSA, and LN, and *ADAPT-*

FORMER introduces non-linear residuals to MLP. However, these approaches do not significantly alter MSA weights, and this limitation can impact their effectiveness when the data distribution undergoes significant changes. In contrast, μ-Tuning employs a hybrid approach, combining linear and non-linear transformations for effective 3D adaptation to result in significantly improved performance.

6.2 The McNemar's Test

To further highlight the performance gains of μ-Tuning compared to other methods, we utilize McNemar's test—a statistical method designed for analyzing paired nominal data [44]. This test is widely used in the machine learning literature to evaluate and compare the predictive accuracy of two models. Figure 4 illustrates this concept, where cell B indicates instances where μ-Tuning as *Model 1* is correct and other baselines as *Model 2* is incorrect, while cell C represents the opposite scenario.

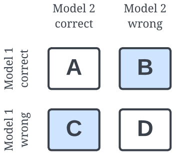

Fig. 4: Here, we take our μ-Tuning as *Model 1* and the compared baseline as *Model 2*.

In Table 1, we present the p-value of McNemar's Test between μ-Tuning as *Model 1* and fine-tuning comparative methods, including BIAS, SSF, ADAPTFORMER, and VPT as *Model 2*. A p-value less than a predefined significance threshold, such as $\alpha = 0.05$, indicates that we can reject the null hypothesis. In other words, it means that the performances of the two models are statistically different. Additionally, the ratio B/C is larger than 1.0 suggesting that the number of correct predictions of μ-Tuning is higher than the compared method. Altogether, the comparisons that have a p-value smaller than 0.05 and a ratio B/C higher than 1.0 indicate our μ-Tuning is superior to the compared method. These results are shaded in cyan color in Table 1. The results suggest that our μ-Tuning is a stronger classifier compared to fine-tuning methods.

6.3 Linear Versus Non-Linear in μ-Tuning

In Fig. 3, a significant difference in performance metrics is observed between linear and non-linear tuning techniques. In 2D tasks, where the input modality closely aligns with that of the pre-trained foundation models, linear tuning methods such as BIAS [8,66] and SSF [37] consistently outperform their non-linear counterparts. Conversely, for 3D tasks, non-linear approaches such as FULL, AdaptFormer [13], and VPT [31] demonstrate superior performances. This observation supports our hypothesis that excessive tuning in 2D tasks disrupts the pre-trained features and subsequently harms performance. On the other hand, inadequate tuning impedes the model's effectiveness in 3D tasks, possibly because the features captured by 2D pre-trained models significantly differ from those required for classifying 3D data. This underscores the importance of a unified approach like μ-Tuning, which automatically adapts the tuning complexity depending on the input data dimensionality.

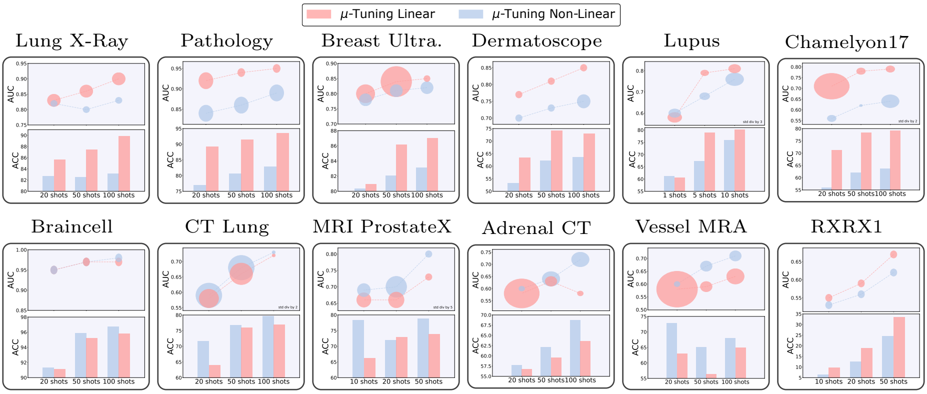

Fig. 5: Comparisons of finetuning results of two variants of our μ-Tuning across many tasks in the few-shot scenarios. The μ-Tuning variants utilize linear and both linear and non-linear tuning approaches, called *μ-Tuning Linear* and *μ-Tuning Non-Linear*, respectively. In the AUC plots, the radius indicates the *standard deviation* of AUC scores from three runs. While *μ-Tuning Linear* surpasses its counterpart in two-dimensional tasks, it is reversed in high-dimensional tasks. The results verify our *μ-Tuning Non-Linear* approach to tuning two- and high-dimensional modalities.

To investigate the impact of non-linear tuning on μ-Tuning, we manipulate the activation of the Adapter blocks in MSA modules in our μ-Tuning framework, forming *μ-Tuning Linear* and *μ-Tuning Non-Linear* versions. In *μ-Tuning Linear*, the Adapter blocks are always inactive, whereas in *μ-Tuning Non-Linear*, they are always activated regardless of the input dimension. We then investigate their efficacy in 2D and 3D tasks. Note that the Adapter block within the initial μ-Tuning is activated only when input dimensionality differs from that of the pre-trained data, as discussed in Section 4 and Fig. 2.

Consistent with the performance patterns observed in the linear and non-linear tuning methodologies, similar trends can be seen in the empirical results of Linear and μ-Tuning Non-Linear, as shown in Figure 5. In the AUC plots, the radii represent the *standard deviation* of AUC scores obtained from three different runs using random seeds. Evidently, *μ-Tuning Linear* excels in 2D tasks but struggles in 3D tasks. In contrast, *μ-Tuning Non-Linear* exhibits an inverse behavior and excels in 3D tasks. This disparity in performance can be attributed to the different nature of their tuning mechanisms. The *μ-Tuning Linear* approach harnesses scaling and shifting transformations via the SSF technique, facilitating linear tuning of the input within the MSA while preserving the pre-trained knowledge. This approach has substantial benefits within the 2D context. Conversely, *μ-Tuning Non-Linear* introduces a bottleneck-like neural network, i.e., the Adapter Block. It functions as a residual connection to the MLP component within the MSA. This architecture allows non-linear adjustments to the pre-trained weights and facilitates more effective adaptation to the intricacies of 3D tasks. These empirical findings validate our initial hypothesis and show the benefits of a unified approach like μ-Tuning.

To provide better insight on μ-Tuning, we introduce further analyses, such as *cross-domain performance*, *stability analysis*, *parameter efficiency*, and *accuracy performance* in *Supplementary Material*.

7 Conclusion

In this study, we conducted a comprehensive benchmarking analysis to evaluate the effectiveness various strategies for fine-tuning a foundation model within medical few-shot scenarios, including visual prompts and minimal weight-tuning techniques. We then introduced a novel method named μ-Tuning, which surpasses popular fine-tuning methods in both stability and performance. Our extensive performance comparison provided new insights into the effectiveness of both linear and non-linear tuning techniques, highlighting their stability and performance across twelve medical imaging tasks. We found that μ-Tuning provides a robust, all-in-one solution for adapting pre-trained foundational models to specific medical imaging applications. We believe this research will significantly enhance the robustness and usability of foundational models in medical contexts.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020) 10
2. Armato III, S.G., Huisman, H., Drukker, K., Hadjiiski, L., Kirby, J.S., Petrick, N., Redmond, G., Giger, M.L., Cha, K., Mamonov, A., et al.: Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. Journal of Medical Imaging **5**(4), 044501–044501 (2018) 10
3. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011) 10
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016) 4
5. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv:2203.17274 (2022) 3
6. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. NIPS **35**, 25005–25017 (2022) 3
7. Bowman, B., Achille, A., Zancato, L., Trager, M., Perera, P., Paolini, G., Soatto, S.: a-la-carte prompt tuning (apt): Combining distinct data via composable prompting. In: CVPR. pp. 14984–14993 (2023) 5
8. Cai, H., Gan, C., Zhu, L., Han, S.: Tinytl: Reduce memory, not parameters for efficient on-device learning. NIPS **33**, 11285–11297 (2020) 2, 3, 5, 7, 11, 12, 13
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017) 10
10. Chavan, A., Liu, Z., Gupta, D., Xing, E., Shen, Z.: One-for-all: Generalized lora for parameter-efficient fine-tuning. arXiv preprint arXiv:2306.07967 (2023) 9
11. Chen, A., Lorenz, P., Yao, Y., Chen, P.Y., Liu, S.: Visual prompting for adversarial robustness. In: ICASSP. pp. 1–5. IEEE (2023) 3

12. Chen, H., Tao, R., Zhang, H., Wang, Y., Ye, W., Wang, J., Hu, G., Savvides, M.: Conv-adapter: Exploring parameter efficient transfer learning for convnets. arXiv:2208.07463 (2022) 4
13. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv:2205.13535 (2022) 2, 4, 5, 6, 7, 9, 11, 12, 13
14. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020) 1
15. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv:2205.08534 (2022) 4
16. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. arXiv:2302.08958 (2023) 3
17. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv:1902.03368 (2019) 10
18. Dong, B., Zhou, P., Yan, S., Zuo, W.: Lpt: Long-tailed prompt tuning for image classification. arXiv:2210.01033 (2022) 3
19. Dong, Y., Cordonnier, J.B., Loukas, A.: Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: ICML. pp. 2793–2803. PMLR (2021) 4, 6, 9
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020) 4, 10
21. Elbatel, M., Martí, R., Li, X.: Fopro-kd: Fourier prompted effective knowledge distillation for long-tailed medical image recognition. arXiv:2305.17421 (2023) 3
22. Feng, C.M., Li, B., Xu, X., Liu, Y., Fu, H., Zuo, W.: Learning federated visual prompt in null space for mri reconstruction. In: CVPR (2023) 3
23. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999) 2
24. He, A., Wang, K., Wang, Z., Li, T., Fu, H.: Dvpt: Dynamic visual prompt tuning of large pre-trained models for medical image analysis (2023) 3, 4
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022) 1
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020) 1
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 4
28. He, X., Yang, D., Feng, W., Fu, T.J., Akula, A., Jampani, V., Narayana, P., Basu, S., Wang, W.Y., Wang, X.E.: Cpl: Counterfactual prompt learning for vision and language models. arXiv:2210.10362 (2022) 3
29. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019) 2, 4, 5, 6, 9, 11
30. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv:2106.09685 (2021) 2, 4, 5, 6, 9, 11
31. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727. Springer (2022) 2, 3, 5, 9, 11, 12, 13

32. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019) 10
33. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018) 10
34. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv:2304.02643 (2023) 1
35. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: ICML. pp. 5637–5664. PMLR (2021) 10
36. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023) 10
37. Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & shifting your features: A new baseline for efficient model tuning. arXiv:2210.08823 (2022) 2, 4, 5, 6, 9, 11, 12, 13
38. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H.: Computer-aided detection of prostate cancer in mri. IEEE TMI **33**(5), 1083–1092 (2014) 10
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 10
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 10
41. Ly, S.T., Lin, B., Vo, H.Q., Maric, D., Roysam, B., Nguyen, H.V.: Student collaboration improves self-supervised learning: Dual-loss adaptive masked autoencoder for brain cell image analysis. arXiv:2205.05194 (2022) 10
42. Maric, D., Jahanipour, J., Li, X.R., Singh, A., Mobiny, A., Van Nguyen, H., Sedlock, A., Grama, K., Roysam, B.: Whole-brain tissue mapping toolkit using large-scale highly multiplexed immunofluorescence imaging and deep neural networks. Nature communications **12**(1), 1550 (2021) 10
43. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989) 2, 5, 11
44. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947) 12, 13
45. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. arXiv:2005.00247 (2020) 5, 9
46. Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: Adapterhub: A framework for adapting transformers. arXiv:2007.07779 (2020) 5, 9
47. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. ArXiv **abs/2209.15517** (2022) 3
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 1, 10, 12
49. Ratcliff, R.: Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological review **97**(2), 285 (1990) 2

50. Saeed, N., Ridzuan, M., Majzoub, R.A., Yaqub, M.: Prompt-based tuning of transformer models for multi-center medical image segmentation of head and neck cancer. Bioengineering **10**(7), 879 (2023) 3
51. Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J.E., Keutzer, K., Darrell, T.: Multitask vision-language prompt tuning. arXiv:2211.11720 (2022) 2, 12
52. Taylor, J., Earnshaw, B., Mabey, B., Victors, M., Yosinski, J.: Rxrx1: An image set for cellular morphological variation across many experimental batches. In: ICLR (2019) 10
53. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018) 10
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS **30** (2017) 5
55. Villa, A., Alcázar, J.L., Alfarra, M., Alhamoud, K., Hurtado, J., Heilbron, F.C., Soto, A., Ghanem, B.: Pivot: Prompting for video continual learning. In: CVPR. pp. 24214–24223 (2023) 3
56. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. arXiv:2203.05962 (2022) 4, 6, 9
57. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: ECCV. pp. 631–648. Springer (2022) 3, 5
58. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: CVPR. pp. 139–149 (2022) 2, 3, 5, 11
59. Wang, Z., Yu, X., Rao, Y., Zhou, J., Lu, J.: P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. NIPS **35**, 14388–14402 (2022) 3
60. Wu, C.H., Motamed, S., Srivastava, S., De la Torre, F.D.: Generative visual prompt: Unifying distributional control of pre-trained generative models. NIPS **35**, 22422–22437 (2022) 3
61. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. NIPS **35**, 124–141 (2022) 3
62. Yang, H., Lin, J., Yang, A., Wang, P., Zhou, C., Yang, H.: Prompt tuning for generative multimodal pretrained models. arXiv:2208.02532 (2022) 3
63. Yang, J., Huang, X., He, Y., Xu, J., Yang, C., Xu, G., Ni, B.: Reinventing 2d convolutions for 3d images. IEEE Journal of Biomedical and Health Informatics **25**(8), 3009–3018 (2021) 10
64. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023) 10
65. Yang, X., Xia, D., Kin, T., Igarashi, T.: Intra: 3d intracranial aneurysm dataset for deep learning. In: CVPR. pp. 2656–2666 (2020) 10
66. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv:2106.10199 (2021) 2, 3, 5, 7, 11, 12, 13
67. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR. pp. 8552–8562 (2022) 3
68. Zhang, X., Iwasawa, Y., Matsuo, Y., Gu, S.S.: Amortized prompt: Lightweight fine-tuning for clip in domain generalization. arXiv:2111.12853 **2** (2021) 3

69. Zhang, Y., Zhou, K., Liu, Z.: Neural prompt search. arXiv:2206.04673 (2022) 9

70. Zhang, Y., Huang, S.C., Zhou, Z., Lungren, M.P., Yeung, S.: Adapting pre-trained vision transformers from 2d to 3d through weight inflation improves medical image segmentation. In: Machine Learning for Health. pp. 391–404. PMLR (2022) 10

71. Zheng, Z., Yue, X., Wang, K., You, Y.: Prompt vision transformer for domain generalization. arXiv:2208.08914 (2022) 3