Personalized Federated Side-Tuning for Medical Image Classification

Jiayi Chen[1,2*], Benteng Ma[1,3*], Yongsheng Pan[1,3], Bin Pu[4], Hengfei Cui[1,3], and Yong Xia[1,3,5(✉)]

[1] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn
[2] Department of Data Science & AI, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia
[3] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[4] Hong Kong University of Science and Technology, Hong Kong SAR, China
[5] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China

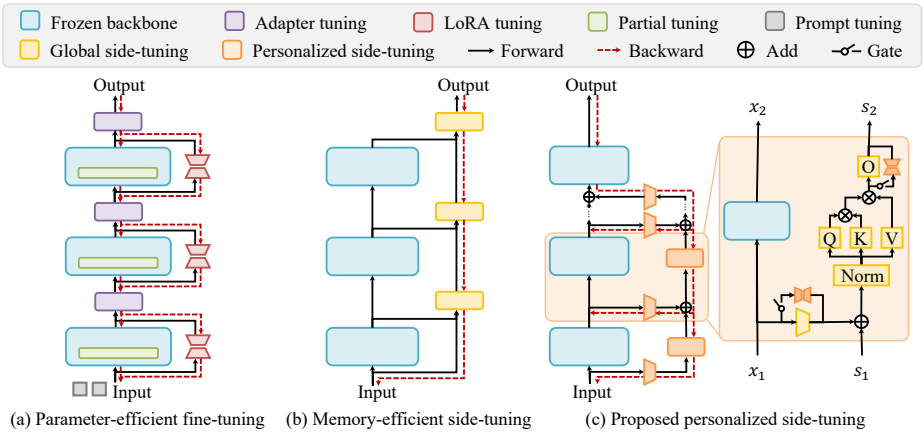
Abstract. Large Vision-Language Models (VLMs) capture rich multimodal knowledge through pretraining and demonstrate remarkable performance across various tasks. However, adapting these foundation models to medical image analysis through fine-tuning faces significant challenges, including constrained computing resources, data privacy concerns, and data heterogeneity. Federated Parameter-Efficient Fine-Tuning (PEFT) emerges as a promising solution, enabling multiple clinical institutions to collaboratively fine-tune VLMs with a small number of parameters. However, it still suffers from data heterogeneity across clients and high training memory requirements. In this work, we propose a **p**ersonalized **Fed**erated **S**ide-**T**uning (pFedST) method. Specifically, we equip each client with a frozen pre-trained CLIP model and a lightweight, learnable, personalized side network for fine-tuning. Only a portion of the side network parameters participates in model aggregation, while the personalized LoRA modules within the side network address data heterogeneity with minimal additional parameters. Extensive experiments demonstrate that pFedST consistently outperforms 12 state-of-the-art methods across two real multi-center medical image classification tasks.

Keywords: Personalized Federated Learning · Side Tuning · Medical Image Classification.

1 Introduction

Large Vision-Language Models (VLMs), such as CLIP [15], have demonstrated remarkable performance and generalization across various tasks by capturing

*J. Chen and B. Ma contributed equally. Corresponding author: Y. Xia.



(a) Parameter-efficient fine-tuning (b) Memory-efficient side-tuning (c) Proposed personalized side-tuning

Fig. 1: Overview of (a) parameter-efficient fine-tuning, (b) memory-efficient side-tuning, and (c) the proposed personalized federated side-tuning (pFedST).

rich multimodal knowledge through pretraining [11]. However, adapting VLMs to medical image analysis [6, 4, 19, 24] faces several significant challenges [29], including strict data privacy regulations [13, 21, 20], limited computational resources, and substantial communication costs across multiple clinical institutions. Federated PEFT [28], which integrates parameter-efficient fine-tuning techniques within federated learning, has emerged as a promising solution. It enables collaborative training of a unified global model without data sharing, while reducing computational resource requirements. However, it often fails to handle the diverse data distributions across local clients, rendering the unified global model less effective in adapting to data heterogeneity [27]. This limitation emphasizes the need for personalized federated PEFT methods that can effectively customize the global model to fit the varied data distributions.

Existing personalized federated PEFT methods can be broadly categorized into three types: prompt-based, adapter-based, and LoRA-based approaches. Prompt-based methods introduce personalized prompts at the input layer to address data heterogeneity, which can be optimized through contrastive learning [10], optimal transport [12], or dynamic prompt generation [23]. Adapter-based methods insert personalized adapters [3, 22] between the frozen blocks of the pretrained model without modifying the pre-trained weights. LoRA-based methods [1] fine-tune the model by inserting low-rank learnable matrices into the frozen layers, enabling personalization with fewer additional parameters. Existing methods face two key challenges: (1) Insufficient personalization. Prompt-based methods in shallow layers cannot capture complex data characteristics, while adapter-based methods require more parameters than LoRA, limiting fine-grained personalization with limited resources. (2) Ineffective computation. As shown in Fig. 1(a), existing PEFT methods still require gradient backpropaga-

tion through the large pre-trained model, resulting in high GPU memory consumption and slow training speed [17].

To deal with the aforementioned two issues, we propose a **p**ersonalized **Fed**erated **S**ide-**T**uning method, termed pFedST. In pFedST, each local client is equipped with a frozen pre-trained CLIP model, consisting of an image encoder and a text encoder. As shown in Fig. 1(c), pFedST introduces a lightweight personalized side network to the image encoder for fine-tuning. This personalized side network offers two significant advantages: (1) pFedST leverages personalized LoRA modules in the side network, tackling data heterogeneity at the cost of minimal additional parameters. (2) pFedST takes the intermediate features of the frozen model as input to the personalized side network, restricting gradient backpropagation to the final frozen block and the lightweight side network. This significantly reduces training memory requirements.

The main contributions are three-fold. (1) We investigate a rarely explored problem, *i.e.* fine-tuning VLMs in federated scenarios considering both data heterogeneity and training memory requirement. (2) We propose a novel personalized federated PEFT method, termed as pFedST, which consists of a personalized side network and a personalized training strategy to tackle data heterogeneity and reduce training memory requirement. (3) pFedST outperforms 12 SOTA methods on two real multi-center medical image classification datasets.

2 Method

As illustrated in Fig. 2, personalized federated learning involves a global model on the server and M clients, each equipped with its own personalized module in addition to the global model. Each client is equipped with a pre-trained CLIP model, comprising an image encoder $\mathcal{I}$ and a text encoder $\mathcal{T}$. In pFedST, we introduce a personalized side network on the image encoder $\mathcal{S}$ for each client (Sec. 2.1), which consists of global side blocks for capturing shared knowledge and local LoRA modules for personalized adaptation. During local fine-tuning, both the global and personalized modules are updated. However, only global side blocks of the side network are aggregated during the aggregation phase.

2.1 Personalized Side Network

As shown in Fig. 2, each client is first equipped with a frozen, pre-trained CLIP model containing an image encoder $\mathcal{I}$ and a text encoder $\mathcal{T}$. The image encoder $\mathcal{I}$ consists of a patch embedding layer followed by L transformer blocks. Let $\boldsymbol{I}_0$ denote the feature output from the patch embedding layer, and $\boldsymbol{I}_i$ denote the feature output from the i-th transformer block, where $i \in \{1, 2, \ldots, L\}$. For the m-th client, a lightweight parallel side network $\mathcal{S}_m$ containing $L-1$ side blocks is injected alongside the image encoder $\mathcal{I}$. For the i-th side block, the inputs $\boldsymbol{X}_i$ consist of two components: the output from the $(i-1)$-th side block $\boldsymbol{S}_{i-1}$ and the projected feature $\boldsymbol{P}_{i-1}$, which is derived from $\boldsymbol{I}_{i-1}$ through a linear down-projection layer. This layer includes a global down-projection matrix

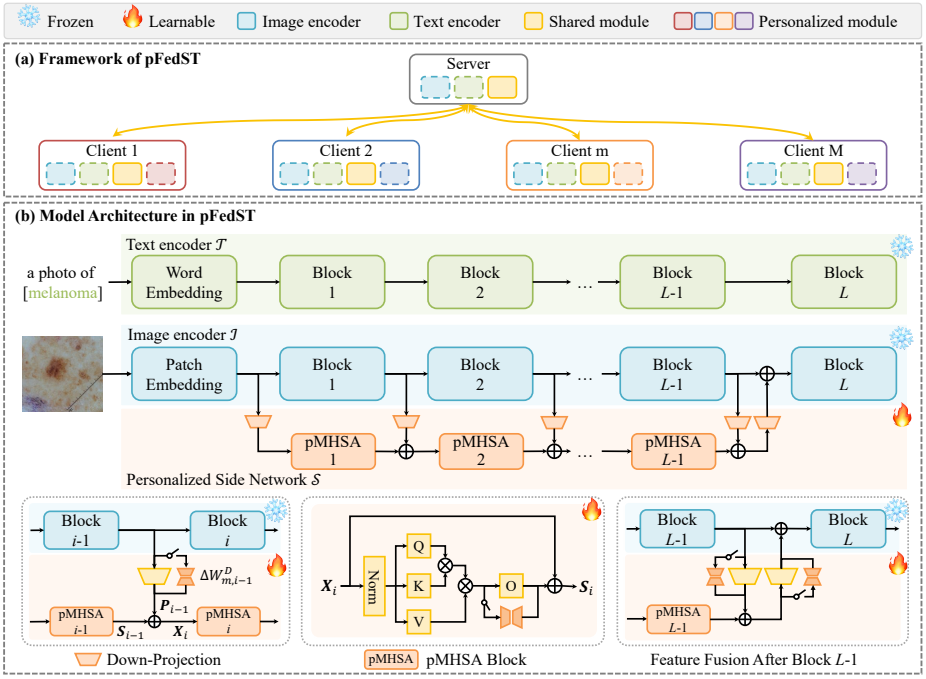


Fig. 2: The framework of pFedST and the architecture of the client model.

$\boldsymbol{W}_i^D \in \mathbb{R}^{d_1 \times d_2}$ and a personalized LoRA module $\Delta\boldsymbol{W}_{m,i}^D \in \mathbb{R}^{d_1 \times d_2}$ for the m-th client. The projected feature $\boldsymbol{P}_i$ is computed as:

$$\boldsymbol{P}_i = LN(\boldsymbol{I}_i(\boldsymbol{W}_i^D + \Delta\boldsymbol{W}_{m,i}^D)) = LN(\boldsymbol{I}_i(\boldsymbol{W}_i^D + \boldsymbol{A}_{m,i}^D\boldsymbol{B}_{m,i}^D)), \tag{1}$$

where $\boldsymbol{A}_{m,i}^D \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{B}_{m,i}^D \in \mathbb{R}^{r \times d_2}$ are low rank projection ($r \ll d_1, r \ll d_2$), and $LN(\cdot)$ denotes the layer norm function. Therefore, the output of the i-th side block is then given by:

$$\boldsymbol{S}_i = f_i(\boldsymbol{X}_i) + \boldsymbol{X}_i = f_i(\boldsymbol{S}_{i-1} + \boldsymbol{P}_{i-1}) + \boldsymbol{S}_{i-1} + \boldsymbol{P}_{i-1}, \quad (\boldsymbol{S}_0 = 0) \tag{2}$$

where $\boldsymbol{S}_i$ denotes the output of the i-th side block, and $f_i(\cdot)$ represents the transformation module that combines the inputs. Specifically, we set $f_i(\cdot)$ to be a personalized multi-head self-attention (pMHSA) as follows:

$$\text{pMHSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \underbrace{\text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)(\boldsymbol{W}_i^O}_{\text{global}} + \underbrace{\Delta\boldsymbol{W}_{m,i}^O}_{\text{personalized}}) \tag{3}$$

where $\boldsymbol{W}_i^O$ is a global linear layer and $\Delta\boldsymbol{W}_{m,i}^O$ is another personalized LoRA module, and each head is:

$$\text{head}_j = \text{Softmax}\left(\frac{(\boldsymbol{X}\boldsymbol{W}_j^Q)(\boldsymbol{X}\boldsymbol{W}_j^K)^\top}{\sqrt{d_j}}\right)(\boldsymbol{X}\boldsymbol{W}_j^V). \tag{4}$$

At the end of the side network, we combine the output of the side network $\boldsymbol{S}_{L-1}$ with the projected feature $\boldsymbol{P}_{L-1}$, and upscale it using a personalized up-projection layer $\boldsymbol{W}_i^U + \Delta\boldsymbol{W}_{m,i}^U$. The upscaled feature is integrated with $\boldsymbol{I}_{L-1}$, and the resulting feature is then fed into the last frozen transformer block $\mathcal{I}_L$ of the image encoder. The final output of the fine-tuned image encoder is

$$\boldsymbol{I} = \mathcal{I}_L(\boldsymbol{I}_{L-1} + LN((\boldsymbol{S}_{L-1} + \boldsymbol{P}_{L-1})(\boldsymbol{W}_{L-1}^U + \Delta\boldsymbol{W}_{m,L-1}^U))), \tag{5}$$

where $\boldsymbol{W}_{L-1}^U$ is a global linear layer, $\Delta\boldsymbol{W}_{m,L-1}^U$ is a personalized LoRA module.

During model aggregation, clients upload only the global modules, including global projections ($\boldsymbol{W}^D$ and $\boldsymbol{W}^U$) and global self-attention blocks ($\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, $\boldsymbol{W}^V$ and $\boldsymbol{W}^O$), while keeping local LoRA modules ($\Delta\boldsymbol{W}^D, \Delta\boldsymbol{W}^U$, and $\Delta\boldsymbol{W}^O$) private for personalization. The server then aggregates these global modules via weighted averaging based on client dataset sizes. The aggregation process is defined as follows:

$$\boldsymbol{W} = \frac{N_m}{\sum_{m=1}^{M} N_m}\boldsymbol{W}_m, \tag{6}$$

where N_m denotes the dataset size of client m, and $\boldsymbol{W} = (\boldsymbol{W}^D, \boldsymbol{W}^U, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{W}^O)$ represents the global modules to be aggregated.

2.2 Training Strategy

Built upon CLIP, pFedST encodes the text feature for class c as $\boldsymbol{T}_c = \mathcal{T}(P; CLS_c)$, where P is a predefined prompt and CLS_c denotes the disease name. Given an input image x, the global image feature $\boldsymbol{I}_G$ is extracted using the global modules. Meanwhile, the personalized image feature $\boldsymbol{I}$ is obtained through the personalized modules. Therefore, the personalized cross-entropy loss is:

$$\mathcal{L}_{ce} = -\frac{1}{N_m}\sum_{n=1}^{N_m} \log \frac{\exp(\cos(\boldsymbol{I}, \boldsymbol{T}_y)/\tau)}{\sum_{c=1}^{C} \exp(\cos(\boldsymbol{I}, \boldsymbol{T}_c)/\tau)}, \tag{7}$$

where $\boldsymbol{T}_y$ is the text feature of ground-truth and τ is the temperature hyperparameter.

Furthermore, we introduce a contrastive loss to preserve shared knowledge by treating the global image and ground-truth text features as positive pairs, while enhancing personalization by treating the global and personalized image features as negative pairs. The contrastive loss is defined as

$$\mathcal{L}_{cont} = -\log \frac{\exp(2\cos(\boldsymbol{I}_G, \boldsymbol{T}^y))}{\exp(2\cos(\boldsymbol{I}_G, \boldsymbol{T}^y)) + \exp(2\cos(\boldsymbol{I}_G, \boldsymbol{I}))}. \tag{8}$$

The overall loss function is

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{cont}. \tag{9}$$

3 Experiments

3.1 Experimental Settings

Datasets and Evaluation Metric. We evaluated the proposed pFedST on two real-world multi-center medical image classification datasets: FedISIC [18] and FedDRG [2]. FedISIC is a skin lesion classification dataset sourced from four distinct medical centers, with sample sizes of 12,413, 3,952, 3,362, and 2,256, respectively [18]. FedDRG is a diabetic retinopathy grading dataset collected from five centers, comprising 3,662, 12,521, 2,000, 516, and 1,744 samples. Each medical center was treated as an independent client, with its data split into training and test sets in an 80%/20% ratio, following official protocols [18, 2]. For performance evaluation, we adopted balanced accuracy (BACC) for FedISIC following [18, 5] and Area Under the Curve (AUC) for FedDRG following [2].
Implemental Details. For both FedISIC and FedDRG datasets, each sample was resized to 256×256 pixels. Training samples were augmented with random rotation, flipping, and random cropping to 224×224 pixels to enhance model robustness and generalization. Test samples were center-cropped to 224×224 pixels for consistency. Federated learning was conducted for T=100 communication rounds, with each client training locally for E=1 epoch per round. The local models were based on the CLIP ViT-B/16 architecture and were optimized using the AdamW optimizer with an initial learning rate of $1e{-}3$, betas set to $(0.9, 0.999)$, and weight decay of $1e{-}2$. The learning rate was adaptively adjusted using a cosine annealing scheduler, decaying to a minimum value of $1e{-}6$ over the training process. Each experiment was repeated three times with different random seeds, and the mean and standard deviation of performance metrics were reported for comprehensive evaluation.
Comparison Methods. We compared pFedST with state-of-the-art methods: (1) **a fully fine-tuned baseline**, Fully-FT; (2) **five federated PEFT methods**—CLIP-Adapter (IJCV24), CLIP-LoRA (CVPRW24), DTL+ (AAAI24), LAST (Arxiv24), and FedTPG (ICLR24); and (3) **six personalized federated PEFT methods**—DPLCLIP (arXiv21), FedTPG* (ICLR24), FedAPT (AAAI24), FedPGP (ICML24), FedOTP (CVPR24), and FedDAT (AAAI24).

3.2 Results

The performance of each client and the average performance across all clients are presented in Table 1 for FedISIC and Table 2 for FedDRG. In these tables, the best results are highlighted in **bold**, while the second-best results are indicated with an underline. As observed from the results, personalized federated PEFT methods consistently outperformed general ones on both datasets, which is attributed to their client-specific adaptation to heterogeneous data distributions. The proposed method, pFedST, demonstrated significant efficacy by consistently achieving the highest average performance across all clients. Specifically, pFedST attained average results of 84.46% on FedISIC and 90.76% on FedDRG, surpassing all competing methods. These results highlight the effectiveness of

Table 1: Performance of pFedST and 12 competing methods on FedISIC.

Method	#Params. (M)	Mem. (GB)	FedISIC (BACC, %) A	B	C	D	Avg
Fully-FT	86.19	5.55	**89.22**$_{\pm 0.66}$	78.15$_{\pm 1.16}$	84.86$_{\pm 1.49}$	62.53$_{\pm 2.38}$	78.69$_{\pm 0.51}$
Federated Parameter-Efficient Fine-Tuning							
CLIP-Adapter [9]	3.67	3.69	85.20$_{\pm 0.81}$	83.55$_{\pm 2.48}$	82.04$_{\pm 2.21}$	61.03$_{\pm 2.71}$	77.98$_{\pm 0.38}$
CLIP-LoRA[25]	**1.18**	5.44	85.20$_{\pm 0.77}$	82.36$_{\pm 1.47}$	83.79$_{\pm 2.52}$	59.96$_{\pm 3.31}$	77.83$_{\pm 1.24}$
DTL+ [8]	1.29	3.02	82.12$_{\pm 0.62}$	88.49$_{\pm 1.79}$	80.47$_{\pm 1.01}$	58.50$_{\pm 2.18}$	77.40$_{\pm 0.50}$
LAST [17]	1.53	**2.69**	82.32$_{\pm 0.31}$	88.95$_{\pm 0.51}$	77.76$_{\pm 0.83}$	61.85$_{\pm 0.70}$	77.72$_{\pm 0.18}$
FedTPG [14]	5.65	3.76	84.67$_{\pm 1.68}$	82.54$_{\pm 3.34}$	84.01$_{\pm 2.27}$	61.79$_{\pm 3.41}$	78.25$_{\pm 1.38}$
Personalzied Federated Parameter-Efficient Fine-Tuning							
DPLCLIP [26]	3.80	3.70	85.72$_{\pm 1.84}$	85.69$_{\pm 3.49}$	83.84$_{\pm 1.22}$	63.23$_{\pm 0.21}$	79.62$_{\pm 1.24}$
FedTPG* [14]	5.65	3.76	83.71$_{\pm 0.39}$	83.45$_{\pm 2.09}$	84.05$_{\pm 0.76}$	65.95$_{\pm 2.66}$	79.29$_{\pm 0.44}$
FedAPT [16]	3.87	3.74	84.47$_{\pm 0.70}$	86.19$_{\pm 0.98}$	83.07$_{\pm 0.64}$	67.19$_{\pm 0.62}$	80.23$_{\pm 0.22}$
FedPGP [7]	3.55	4.00	84.85$_{\pm 1.18}$	88.75$_{\pm 2.09}$	85.80$_{\pm 1.34}$	71.98$_{\pm 0.22}$	82.84$_{\pm 0.86}$
FedOTP [12]	3.56	4.05	85.61$_{\pm 0.57}$	87.92$_{\pm 4.05}$	83.90$_{\pm 2.08}$	68.14$_{\pm 2.01}$	81.39$_{\pm 2.04}$
FedDAT [3]	2.59	3.69	86.53$_{\pm 0.81}$	84.83$_{\pm 1.04}$	86.25$_{\pm 1.31}$	74.90$_{\pm 0.59}$	83.13$_{\pm 0.35}$
pFedST (Ours)	2.32	2.98	84.50$_{\pm 0.57}$	**89.52**$_{\pm 1.65}$	**88.70**$_{\pm 0.43}$	**75.13**$_{\pm 0.13}$	**84.46**$_{\pm 0.35}$

Table 2: Performance of pFedST and 12 competing methods on FedDRG.

Method	#Params (M)	Mem. (GB)	FedDRG (AUC, %) A	B	C	D	E	Avg
Fully-FT	86.19	5.55	87.14$_{\pm 0.76}$	90.81$_{\pm 0.88}$	87.78$_{\pm 0.42}$	89.64$_{\pm 0.50}$	82.44$_{\pm 0.76}$	87.56$_{\pm 0.33}$
Federated Parameter-Efficient Fine-Tuning								
CLIP-Adapter [9]	3.67	3.69	89.12$_{\pm 0.37}$	88.99$_{\pm 0.27}$	86.59$_{\pm 2.32}$	88.51$_{\pm 1.45}$	81.39$_{\pm 0.94}$	86.92$_{\pm 0.86}$
CLIP-LoRA[25]	**1.18**	5.44	88.33$_{\pm 0.23}$	89.78$_{\pm 0.44}$	85.95$_{\pm 1.72}$	90.15$_{\pm 0.56}$	81.67$_{\pm 0.51}$	87.17$_{\pm 0.09}$
DTL+ [8]	1.29	3.02	87.45$_{\pm 0.51}$	90.28$_{\pm 0.79}$	87.82$_{\pm 0.55}$	89.34$_{\pm 1.06}$	81.52$_{\pm 0.77}$	87.28$_{\pm 0.25}$
LAST [17]	1.53	2.69	85.94$_{\pm 0.95}$	90.29$_{\pm 0.16}$	88.06$_{\pm 0.38}$	89.25$_{\pm 0.75}$	82.02$_{\pm 0.12}$	87.11$_{\pm 0.18}$
FedTPG [14]	5.65	3.76	89.01$_{\pm 0.83}$	89.42$_{\pm 0.35}$	88.35$_{\pm 0.84}$	88.79$_{\pm 0.83}$	80.70$_{\pm 0.53}$	87.25$_{\pm 0.17}$
Personalized Federated Parameter-Efficient Fine-Tuning								
DPLCLIP [26]	3.80	3.70	89.66$_{\pm 0.92}$	87.52$_{\pm 0.17}$	88.84$_{\pm 0.78}$	90.48$_{\pm 0.51}$	84.93$_{\pm 0.85}$	88.27$_{\pm 0.30}$
FedTPG* [14]	5.65	3.76	90.53$_{\pm 0.30}$	88.86$_{\pm 0.59}$	91.92$_{\pm 0.36}$	91.78$_{\pm 0.79}$	**84.89**$_{\pm 1.20}$	89.60$_{\pm 0.58}$
FedAPT [16]	3.87	3.74	91.02$_{\pm 0.21}$	89.47$_{\pm 0.05}$	91.34$_{\pm 1.03}$	92.49$_{\pm 0.16}$	83.40$_{\pm 1.44}$	89.54$_{\pm 0.33}$
FedPGP [7]	3.55	4.00	90.63$_{\pm 1.01}$	89.38$_{\pm 1.21}$	91.14$_{\pm 0.44}$	92.63$_{\pm 0.75}$	83.33$_{\pm 0.35}$	89.42$_{\pm 0.27}$
FedOTP [12]	3.56	4.05	90.81$_{\pm 0.17}$	90.12$_{\pm 0.05}$	92.43$_{\pm 0.22}$	93.17$_{\pm 0.71}$	84.22$_{\pm 0.37}$	90.15$_{\pm 0.06}$
FedDAT [3]	2.59	3.69	**91.21**$_{\pm 0.22}$	91.36$_{\pm 0.22}$	93.57$_{\pm 0.09}$	90.95$_{\pm 0.44}$	83.87$_{\pm 0.28}$	90.19$_{\pm 0.17}$
pFedST (Ours)	2.32	2.98	90.41$_{\pm 0.05}$	**91.40**$_{\pm 0.28}$	**93.83**$_{\pm 0.73}$	**93.81**$_{\pm 0.17}$	84.35$_{\pm 0.12}$	**90.76**$_{\pm 0.21}$

pFedST in addressing the challenges posed by data heterogeneity in federated learning settings. For the FedISIC dataset, pFedST improved balanced accuracy on clients B, C, and D by 0.57%, 2.45%, and 0.23%, respectively, over the second-best method. Additionally, pFedST increased the average performance by 5.77% over the fully fine-tuned baseline and by 1.36% over FedDAT. For the FedDRG dataset, pFedST demonstrated a 3.20% increase in average AUC over fully fine-tuning and exceeded the second-best method by 0.57%. In terms of computational efficiency, pFedST requires only 2.32 million training parameters and 2.98 GB of GPU memory to fine-tune a model based on CLIP ViT/B-16. This represents a reduction of 97.30% in the number of training parameters and a 46.31% decrease in memory usage compared to fully fine-tuning. These findings underscore the ability of pFedST to achieve state-of-the-art performance while significantly reducing computational overhead.

Table 3: Effect of personalized LoRA placement. 'Proj' includes both down-projection and up-projection.

Personalized Weight					FedISIC (BACC, %)				
Proj	Q	K	V	O	A	B	C	D	Avg
					84.64	89.22	81.77	58.94	78.64
✓					82.80	88.33	83.54	67.92	80.65
✓	✓				84.07	85.42	86.45	70.19	81.53
✓		✓			**84.96**	84.78	85.19	71.06	81.50
✓			✓		84.47	**91.16**	86.29	72.56	83.62
✓				✓	84.50	89.52	**88.70**	**75.13**	**84.46**
✓	✓	✓			84.39	95.50	85.58	69.66	83.78
✓	✓		✓		81.95	89.04	85.25	71.46	81.93
✓	✓	✓	✓	✓	83.69	85.55	85.06	69.87	81.04

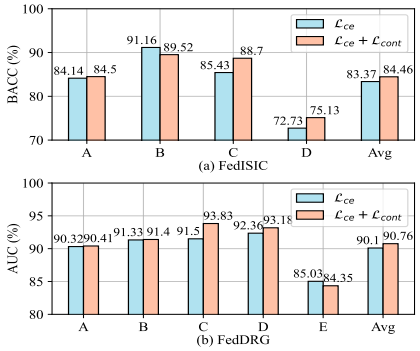


Fig. 3: Effect of contrastive loss

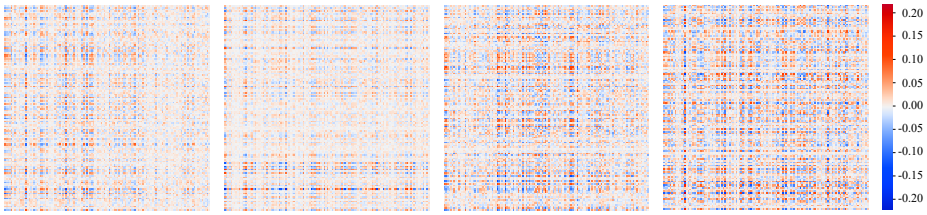


Fig. 4: Visualization of personalized LoRA parameters on FedISIC dataset

3.3 Ablation Study and Further Analysis

Effect of Personalized LoRA Adaptation. As shown in Table 3, we investigate the effects of applying personalized LoRA at different locations within the model on FedISIC. Without personalized LoRA, the baseline model (row 1) achieved an average BACC of 78.64%. In contrast, introducing personalized LoRA to the down-projection and up-projection weights (row 2) resulted in a 2.01% improvement in performance. Notably, personalizing the output projection weight W^O with LoRA achieved the highest performance.

Effect of Contrastive Loss. We evaluated the impact of contrastive loss on FedISIC and FedDRG datasets, as shown in Fig. 3. The results demonstrate that introducing contrastive loss significantly improves both the average performance across clients and the performance of most individual clients.

Visualization of Personalized LoRA. As shown in Fig. 4, we visualized the personalized LoRA parameters for the output projection weight in the first self-attention module on FedISIC. Clients C and D exhibited more concentrated LoRA parameters than A and B, indicating that personalized LoRA effectively captures data heterogeneity and enhances personalization on complex datasets. Table 3 shows that adding personalized LoRA improved BACC by 1.70%, 1.19%, 5.16%, and 16.19% for clients A, B, C, and D, respectively. This highlights how concentrated LoRA parameters adapt to diverse data and boost personalization.

4 Conclusion

We propose pFedST to address data heterogeneity and memory inefficiency in personalized federated PEFT across multiple medical centers. Built upon the large vision-language model CLIP, pFedST introduces a personalized side network for each client and employs a parameter-efficient fine-tuning strategy to adapt the frozen CLIP model to the local client data. This enables more effective modeling of client-specific characteristics. Extensive experiments demonstrate the consistent superiority of pFedST across two real-world multi-center medical image classification tasks.

Acknowledgments. This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), in part by the National Natural Science Foundation of China under Grants 62271405, 6240012686, 62171377, and 92470101, in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), in part by Ningbo Leading Medical&Health Discipline under Grants 2022-S02, in part by Key Research and Development Program of Shaanxi under Grant 2025CY-YBXM-039, and in part by the Shenzhen Science and Technology Program under Grants JCYJ20220530161616036.

Disclosure of Interests. The authors declare no relevant competing interests.

References

1. Babakniya, S., Elkordy, A.R., Ezzeldin, Y.H., Liu, Q., Song, K.B., El-Khamy, M., Avestimehr, S.: SLoRA: Federated parameter efficient fine-tuning of language models. arXiv preprint arXiv:2308.06522 (2023)
2. Che, H., Cheng, Y., Jin, H., Chen, H.: Towards generalizable diabetic retinopathy grading in unseen domains. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 430–440. Springer (2023)
3. Chen, H., Zhang, Y., Krompass, D., Gu, J., Tresp, V.: FedDAT: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 11285–11293 (2024)
4. Chen, J., Ma, B., Cui, H., Xia, Y.: FedEvi: Improving federated medical image segmentation via evidential weight aggregation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 361–372. Springer (2024)
5. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11439–11449 (2024)
6. Chen, J., Ma, B., Cui, H., Zhang, J., Xia, Y.: Active learning based on temporal difference of gradient flow in thoracic disease diagnosis. IEEE Journal of Biomedical and Health Informatics (2025)
7. Cui, T., Li, H., Wang, J., Shi, Y.: Harmonizing generalization and personalization in federated prompt learning. In: Forty-first International Conference on Machine Learning (2024)

8. Fu, M., Zhu, K., Wu, J.: DTL: Disentangled transfer learning for visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 12082–12090 (2024)
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
10. Guo, T., Guo, S., Wang, J.: pFedPrompt: Learning personalized prompt for vision-language models in federated learning. In: Proceedings of the ACM Web Conference 2023. pp. 1364–1374 (2023)
11. Kang, Y., Fan, T., Gu, H., Zhang, X., Fan, L., Yang, Q.: Grounding foundation models through federated transfer learning: A general framework. arXiv preprint arXiv:2311.17431 (2023)
12. Li, H., Huang, W., Wang, J., Shi, Y.: Global and local prompts cooperation via optimal transport for federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12151–12161 (2024)
13. Ma, B., Feng, Y., Chen, G., Li, C., Xia, Y.: Federated adaptive reweighting for medical image classification. Pattern Recognition **144**, 109880 (2023)
14. Qiu, C., Li, X., Mummadi, C.K., Ganesh, M.R., Li, Z., Peng, L., Lin, W.Y.: Federated text-driven prompt generation for vision-language models. In: The Twelfth International Conference on Learning Representations (2024)
15. 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: International conference on machine learning. pp. 8748–8763. PMLR (2021)
16. Su, S., Yang, M., Li, B., Xue, X.: Federated adaptive prompt tuning for multi-domain collaborative learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15117–15125 (2024)
17. Tang, N., Fu, M., Zhu, K., Wu, J.: Low-rank attention side-tuning for parameter-efficient fine-tuning. arXiv preprint arXiv:2402.04009 (2024)
18. Ogier du Terrail, J., Ayed, S.S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al.: Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. Advances in Neural Information Processing Systems **35**, 5315–5334 (2022)
19. Wu, Y., Wu, Z., Shi, H., Picker, B., Chong, W., Cai, J.: Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 3–13. Springer (2023)
20. Wu, Z., Wu, Y., Lin, G., Cai, J.: Reliability-adaptive consistency regularization for weakly-supervised point cloud segmentation. International Journal of Computer Vision **132**(6), 2276–2289 (2024)
21. Xia, Y., Ma, B., Dou, Q., Xia, Y.: Enhancing federated learning performance fairness via collaboration graph-based reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–272. Springer (2024)
22. Xie, C., Huang, D.A., Chu, W., Xu, D., Xiao, C., Li, B., Anandkumar, A.: PerAda: Parameter-efficient federated learning personalization with generalization guarantees. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23838–23848 (2024)
23. Yang, F.E., Wang, C.Y., Wang, Y.C.F.: Efficient model personalization in federated learning via client-specific prompt generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19159–19168 (2023)

24. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: 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. pp. 11114–11124 (2024)
25. Zanella, M., Ben Ayed, I.: Low-rank few-shot adaptation of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1593–1603 (2024)
26. Zhang, X., Gu, S.S., Matsuo, Y., Iwasawa, Y.: Domain prompt learning for efficiently adapting CLIP to unseen domains. Transactions of the Japanese Society for Artificial Intelligence **38**(6), B–MC2_1 (2023)
27. Zhang, Y., Qin, Z., Wu, Z., Deng, S.: Personalized federated fine-tuning for llms via data-driven heterogeneous model architectures. arXiv preprint arXiv:2411.19128 (2024)
28. Zhang, Z., Yang, Y., Dai, Y., Wang, Q., Yu, Y., Qu, L., Xu, Z.: FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In: Annual Meeting of the Association of Computational Linguistics 2023. pp. 9963–9977. Association for Computational Linguistics (ACL) (2023)
29. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: CLIP in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)