Restyled, Tuning, and Alignment: Taming VLMs for Federated Non-IID Medical Image Analysis

Shengchao Chen[1,3,†][0000−0001−9992−2264] and Ting Shu[1,2,‡][0000−0002−8630−7868]

[1] School of Artificial Intelligence, Shenzhen University, Shenzhen 518060, China
[2] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
[3] Australian AI Institute, University of Technology, NSW 2007, Australia
shengchao.chen.uts@gmail.com, tingshu@szu.edu.cn

Abstract. Adapting pretrained Vision Language Models like CLIP, for medical image analysis in federated learning (FL) offers cross-modal insights while preserving privacy. However, effective cross-domain federated adaptation requires intensive fine-tuning and knowledge sharing, challenging in low-resource medical practice due to the divergence between pretrained natural image and medical imagery. Moreover, the significant statistical heterogeneity (non-IID) of medical data exacerbates these challenges. To address these issues, this paper introduces a framework that tames CLIP for non-IID federated medical image classification. This develops client-specific personalized models by reinforcement and constrain local cross-modal alignment, enabling the models to integrate client-specific and globally common knowledge. This approach not only addresses non-IID challenges but also optimizes the trade-off between performance and efficiency. Extensive experiments on real-world medical image datasets confirm the effectiveness and superiority of our FedTCA.

Keywords: Federated Vision-language Models · Medical Image Classification · Prompting · Cross-domain Adaption · Cross-modal Alignment

1 Introduction

Advances in deep medical image classification models rely heavily on extensive training data [5]. However, the availability of such data is often constrained, and medical organizations hesitate to share sensitive information due to privacy laws [11], making centralized training impractical. Furthermore, the limited data access within each organization impedes the acquisition of broader knowledge from exclusive use of their proprietary datasets. Federated Learning (FL) [20] has emerged as a promising learning paradigm, allowing multiple organizations to collaboratively train models while ensuring privacy and eliminating data silos.

Statistical heterogeneity (non-IID) in medical image data across organizations complicates the development of robust models through vanilla FL, hindering performance consistency [3]. Personalized FL (PFL) addresses this by

† Work done during a project internship
‡ Corresponding author

developing client-specific models that integrate local insights with global knowledge to enhance performance [26]. One common method involves adding constraint terms, such as Moreau envelopes [26] or meta-learning techniques [10], to optimization objectives to foster personalization. However, these methods often overlook the model's sensitivity to non-IID data. To counter this, some strategies improve personalization by localizing parameters sensitive to heterogeneous data, such as classifiers [1, 25, 30] or normalization layers [16]. Nonetheless, skewed data distributions can cause these models to overemphasize simple patterns and neglect complex minority data, increasing local bias and diminishing generalizability. Moreover, these approaches often require the transmission of numerous model parameters to facilitate knowledge exchange, and developing and transmitting large-scale models from scratch—which are essential for optimal performance—is impractical in resource-constrained healthcare environments.

Using advanced pretrained models like the vision-language model CLIP [23], pretrained on extensive multimodal data, can bypass the need for costly training from scratch [2, 23]. CLIP, which maps images and texts into a unified representational space, potentially lowers computational costs and deepens insights in federated medical image analysis [2, 27, 17]. However, empirical analyses [15] suggest that CLIP often falls short in medical image classification due to the significant mismatch between training on natural images and applying to medical images, impairing effective cross-domain knowledge transfer [15]. While extensive fine-tuning might address this, it requires substantial resources. Moreover, the complexity of non-IID data in FL environments complicates adaptation. Although FACMIC [27] introduces a lightweight feature attention module and adaptive loss to balance global and local features, it still grapples with trade-offs between global knowledge and local personalization, potentially introducing decision biases. Additionally, communication overhead remains a crucial challenge. This raises an essential research question: *Can we reduce the bias between global and personalized knowledge while maintaining efficiency in adapting the pretrained CLIP to federated medical image classification with non-IID data?*

This paper presents FedTCA, designed to address the research problem outlined above. We introduce a Prompt Restyling strategy that replaces traditional hand-crafted prompts with informative, learnable, and domain-specific alternatives. This strategy also incorporates a low-rank adaptation to enhance pretrained models' capacity for cross-domain knowledge transfer. Additionally, we propose the Twin Cross-modal Alignment (TCA) to mitigate learning biases from non-IID data and to optimize the balance between global and personalized knowledge. TCA improves local visual-text alignment by integrating personalized and global insights and conceptualizing global-local knowledge transfer as an optimal transport problem for efficient resolution. To reduce communication overhead, clients transmit only a subset of parameters. Key contributions are:

- FedTCA, a framework that adapts pretrained CLIP to federated medical image classification with non-IID data while keeping efficiency.
- Prompt Restyling, a method refines prompts with informative, context-aware, and domain-specific alternatives for better cross-domain adaptation.

– Twin Cross-domain Alignment, a approach achieves forward-looking visual-text alignment by effectively integrating personalization and global insights.
– Extensive experiments on four real-world medical imaging datasets demonstrate the effectiveness and superiority of FedTCA.

2 Methodology

2.1 Problem Definition

Considering a FL framework involving multiple medical organizations, each holding a labeled medical images dataset $(x, y) \in \mathcal{D}_i$ across n_i classes. This framework includes a server for model aggregation. Due to non-IID data distributions among clients, where $\mathcal{P}_i(y) \neq \mathcal{P}_j(y)$ yet $\mathcal{P}_i(x|y) = \mathcal{P}_j(x|y)$.To tackle this challenge, we aim to develop personalized models that integrate both client-specific knowledge and shared insights, utilizing the capabilities of pretrained CLIP. Although clients can bypass the huge burden of model initialization with large medical datasets by using pretrained models, cross-domain knowledge transfer (natural to medical images) continues to pose significant challenges on local computational resources, global communication overhead and privacy.

2.2 Framework: FedTCA

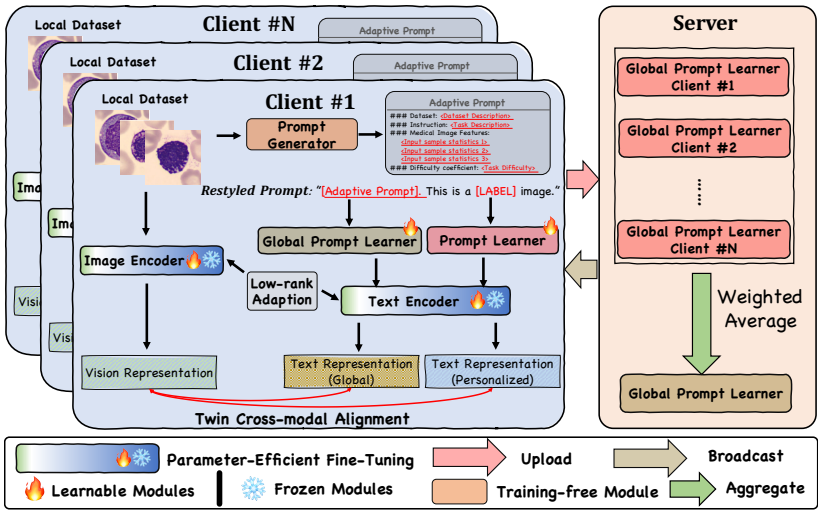

Fig. 1. Schematic diagram of our proposed FedTCA framework. Note that the global prompt learner on the client is identical to the local one upon framework initialization.

Fig. 1 shwos our framework, where clients employ a pretrained CLIP model consisting of image encoder $\boldsymbol{E}_{\text{image}}$ and text encoder $\boldsymbol{E}_{\text{text}}$. During local updates, we restyle the original prompt with a generator, use Prompt Learner

for context-aware embeddings, fine-tune $\boldsymbol{E}_{\text{image}}$ and $\boldsymbol{E}_{\text{text}}$ with low-rank adaption (LoRA) [14] for rapid adaptation, and apply optimal transport-based Twin Cross-modal Alignment for cross-modal alignment. Clients globalize only the Prompt Learner, keeping other parameters local to ensure personalization and privacy. Next, we will introduce these components in details.

Prompt Restyling. Prompting effectively activates task-specific capabilities in VLMs, but CLIP's hand-crafted prompts '`A photo of a [CLS]`' not friendly to medical images because they lacked targeted medical knowledge and sensitivity to detail, resulting in an inability to effectively transfer cross-domain knowledge and align multi-modal representation. To address this, we propose a Prompt Generator to restyle the original prompt by integrating four aspects of domain-specific knowledge: `[Dataset Description]`, `[Task Description]`, `[Input Sample Statistics]`, and `[Task Difficulty]`. A example is as below:

> This dataset consists of microscopic peripheral blood cell images. Predict the label given the input sample from 3-class including eosinophil, erythroblast and platelet. The input sample has a minimum of 17.21, a maximum of 124.1, and a median of 59. The gray-level covariance matrix are $\boldsymbol{M}$. The task is easy.

The restyled prompt refines medical-specific knowledge and enhances pattern recognition for text encoding. Subsequently, this prompt, denoted as Ω_r, replaces the original to initialize the Global Prompt Learner $\mathcal{P}_g$ and the Personalized Prompt Learner $\mathcal{P}_p$ via $\mathcal{P}_g(\Omega_r)$ and $\mathcal{P}_p(\Omega_r)$. Our Prompt Learner distinguishes itself from CoOp's [31, 32], which adds a learnable vector at specific positions, by incorporating learnable parameters for each token, thereby enhancing context-awareness and facilitating cross-modal alignment in medical knowledge.

Local Updating. We used low-rank adaption (LoRA) [14] to improve cross-domain knowledge transfer while keeping efficiency. It adapts the *Query* and *Value* of attention blocks in both $\boldsymbol{E}_{\text{image}}$ and $\boldsymbol{E}_{\text{text}}$, generating low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$ from pretrained weights $\boldsymbol{W} \in \mathbb{R}^{d\times d}$. These matrices transform middle image representations via $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{X} + AB$, with only A, B being trainable, optimizing with the number of dimension d and rank value r, where $r \ll d$. In LoRA, A is randomly initialized and B is initialized to zero.

Twin Cross-modal Alignment. Non-IID data across clients can leads to a learning bias in personalized models where simpler majority representations are favored over more complex minority ones. To achieve a balance between global and personalized knowledge, we propose Twin Cross-modal Alignment (TCA) to strengthen the collaboration between global and local prompt learners based on Optimal Transport (OT) [21], effectively addressing both label shift and feature shift data heterogeneity. Firstly, we define the global and personalized text representation as $\boldsymbol{T}_g \in \mathbb{R}^{L\times d}$ and $\boldsymbol{T}_p \in \mathbb{R}^{L\times d}$ with the corresponding combination $\boldsymbol{T} = [\boldsymbol{T}_g, \boldsymbol{T}_p] \in \mathbb{R}^{d\times 2}$. We consider learning an OT plan $\mathcal{T}$ that aligns both global and local text representation $\boldsymbol{T}$ with vision representation $\boldsymbol{X} \in \mathbb{R}^{V\times d}$. By representing features as samples from discrete distributions, the cost matrix can be

represented by the cosine distance between $\boldsymbol{T}$ and $\boldsymbol{I}$ as $\boldsymbol{C} = 1 - [\boldsymbol{X}^{\mathrm{T}}\boldsymbol{T}] \in \mathbb{R}^{V\times 2}$, then the optimization objective of the optimal transport is formulated as:

$$d_{\boldsymbol{C}}(\alpha,\beta) = \min_{\mathcal{T}\in U(\alpha,\beta)} < \boldsymbol{C}, \mathcal{T} >, U(\alpha,\beta) = \{\mathcal{T} \in \mathbb{R}_{+}^{V\times 2} \mid \mathcal{T}\mathbb{1}_2 < \alpha, \mathcal{T}^{\top}\mathbb{1}_V = \beta\} \quad (1)$$

where $< \cdot, \cdot >$ is Frobenius dot-product, $U(\alpha,\beta)$ is the solution space of $\mathcal{T}$ with $\alpha \in \mathbb{R}^V, \beta \in \mathbb{R}^2$ that are essentially marginal probability vectors which satisfy $||\alpha||_1 \geq ||\beta||_1 = \gamma$ ($\gamma \in [0,1]$). The difference between Eq. 1 and formulation in PLOT lies in their use of classical OT with two hard equality constraints as Eq. (3). This forces prompts to map toeach image patch, potentially causing them to capture someclass-irrelevant information from the image and thereby influencing the final results. In contrast, our method relaxes one of the equality constraints, allowing prompts to concentrate solely on the most relevant image patches rather than the entire content of the image. Additionally, by controlling γ, our method owns the ability to regulate the mapping size of prompts on the feature map. In addition, for fast optimization, we add an entropic regularization term following Sinkhorn algorithm [7] to achieve lightspeed computation as:

$$d_C(\alpha,\beta) = \min_{\mathcal{T}\in U(\alpha,\beta)} < \boldsymbol{C}, \mathcal{T} > +\eta < \mathcal{T}, \log \mathcal{T} >, \quad (2)$$

where $\eta > 0$ is a hyperparameter. We further reformulate Eq. 2 as a Kullback-Leibler (KL) projection with an exponential reference distribution $e^{-\boldsymbol{C}/\eta}$ to explicitly minimize KL divergence as the optimization objective. The solution space $U(\alpha,\beta)$ is defined as the intersection of two convex, non-affine sets as:

$$\begin{gathered} d_{\boldsymbol{C}}(\alpha,\beta) = \min_{\mathcal{T}\in U(\alpha,\beta)} \eta D_{\mathrm{KL}}(\mathcal{T}||e^{-\boldsymbol{C}/\eta}), \\ \boldsymbol{C}_1 \triangleq \{\mathcal{T} \in \mathbb{R}_{+}^{V\times 2} \mid \mathcal{T}\mathbb{1}_2 \leq \alpha\}, \quad \boldsymbol{C}_2 \triangleq \{\mathcal{T} \in \mathbb{R}_{+}^{V\times 2} \mid \mathcal{T}^{\top}\mathbb{1}_V = \beta\}. \end{gathered} \quad (3)$$

We employ a rapid implementation of Dykstra's algorithm [9], which effectively scales iterative KL projection between $\mathcal{C}_1$ and $\mathcal{C}_2$ by leveraging matrix-vector multiplications exclusively. Initializing $Q = \exp(-C/\eta)$ and $v^{(0)} = \mathbb{1}_2$, a fast optimization solution is achieved within a few iterations as $T^* = \mathrm{diag}(u^{(\hat{t})})Q\mathrm{diag}(v^{(\hat{t})})$, where $\hat{t}$ is the iteration, and in each iteration $u^{(\hat{t})} = \min(\mathbb{1}_V/Q_\alpha, \mathbb{1}_V)$ and $v^{(\hat{t})} = \mathbb{1}_2/Q_\beta^T u^{(\hat{t})}$ with $Q_\alpha = Q/\mathrm{diag}(\alpha)\mathbb{1}_{V\times 2}$ and $Q_\beta^T = Q^T/\mathrm{diag}(\beta)\mathbb{1}_{V\times 2}$. Therefore, we can get the optimal transplort plan T^* and the final Wassertein distance d_C, then the twin cross-modal alignment score can be formulated as:

$$\mathcal{L}_{\mathbf{TCA}} = \frac{exp(cos(\boldsymbol{E}_{\mathrm{image}}(\boldsymbol{X}), \boldsymbol{E}_{\mathrm{text}}(\boldsymbol{T}))/\tau)}{\sum_i^n exp(cos(\boldsymbol{E}_{\mathrm{image}}(\boldsymbol{X})), \boldsymbol{E}_{\mathrm{text}}(\boldsymbol{T})/\tau)} \Rightarrow \frac{exp((1-d_{\boldsymbol{C},k})/\tau)}{\sum_i^n exp((1-d_{\boldsymbol{C},c}))/\tau}. \quad (4)$$

After this, we fix the plan T^* and optimized prompt learner in both global and local simultaneously for a specific client through standard cross-entropy loss.

Communication. During communication phase, only Global Prompt Learner $\mathcal{P}_g$ is uploaded, while other parameters (i.e., Local Prompt Learner $\mathcal{P}_p$ and low-rank matrix A/B) remain locally to ensure personalization and privacy. The

server updates $\mathcal{P}_g$ based on FedAvg: $\mathcal{P}_g^{t+1} = \sum_{i \in C_t} \kappa \mathcal{P}_{g,i}^t$, where κ is the sample ratio from the i-th client to the total, and t denotes the round. This updates the Global Prompt Learner for subsequent training rounds.

3 Experiment and Results

Datasets. We used four publicly available medical imaging datasets [28]: BloodMNIST, TiuuseMNIST, and OrganMNIST 2D and OrganMNIST 3D, each with images of 224×224. Considering the original dimensions of the 3D dataset are 64x64x64, we resized each two-dimensional slice (64×64) to 224×224, resulting in dimensions of $224 \times 224 \times 64$ to match the model input requirements. We divided the images into 20 clients (10 clients for OrganMNIST 3D) according Dirichlet distribution as $\texttt{Dirichlet}(\lambda \mathbf{p})$, where $\mathbf{p}$ is the prior class distribution and λ adjusts the non-IID severity, the lower λ, the higher the degree of non-IID. We evaluated three values of $\lambda \in \{0.1, 0.3, 0.5\}$ to assess performance across various non-IID scenarios. Dataset details and distribution are shown in **Fig.2**.

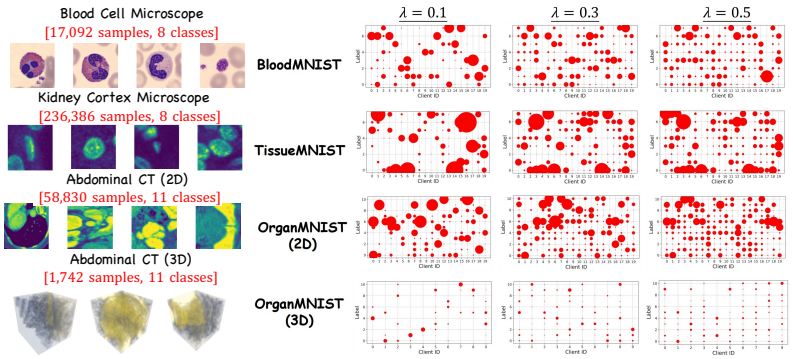

Fig. 2. Datasets. **(Left)** partial sample visualization; **(Right)** distribution in FL setup.

Implementation. We use the `ViT-B/16` [8] as the image encoder for both non-CLIP baseline and CLIP-based baseline. All were trained over 100 communication rounds, each consisting of a single local epoch for fast adaption, with 50% join ratio and SGD at a learning rate of $1e^{-4}$. By default, the rank in LoRA is set to 4 , and batch size is 32. For task difficulty in Prompt Restyling, we assign `hard/medium/easy` corresponding to $\lambda = 0.1/0.3/0.5$. The results were evaluated based on Top-1 accuracy, averaged over five trials, with standard deviations reported. All algorithms are implemented on an Nvidia A2 (16GB) GPU.

Compare with SOTA FL. We compared with FL baselines: FedAvg [20], FedBN [16], PerFedAvg [10], FedALA [29], FedCLIP [18], and FACMIC [27]. **Table 1** demonstrates that our FedTCA consistently outperforms both general

approaches to the Non-IID problem (e.g., FedBN, PerFedAvg) and specialized methods for CLIP adaptation in medical image classification (FACMIC). Our FedTCA improves upon the previous SOTA FACMIC by margins ranging from 5.13% to 15.10% across various datasets and Non-IID setups. These improvements are attributed to our enhancement of local cross-modal alignment through flexible context-aware prompts and a twin cross-modal alignment strategy, which effectively balances global and personalized knowledge to minimize learning bias. Additionally, **Figure 3** illustrates that our FedTCA facilitates the formation of distinct, recognizable clusters, crucial for effective classification.

Table 1. Main results across different non-IID setups. **Bold**: the best.

Dataset	BloodMNIST			TissueMNIST			OrganMNIST (2D)			OrganMNIST (3D)		
Method	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
FedAvg	$63.04_{1.24}$	$68.23_{1.00}$	$71.55_{1.12}$	$60.65_{1.53}$	$65.42_{1.87}$	$66.49_{0.78}$	$69.96_{1.24}$	$74.93_{0.12}$	$79.32_{0.12}$	$62.34_{0.28}$	$68.37_{0.55}$	$71.20_{1.02}$
Local	$84.47_{1.43}$	$86.81_{1.20}$	$87.02_{0.55}$	$73.42_{0.95}$	$77.03_{1.32}$	$79.24_{1.43}$	$81.84_{0.45}$	$87.93_{0.53}$	$91.06_{1.28}$	$88.73_{0.44}$	$91.24_{0.96}$	$92.31_{0.32}$
FedBN	$79.55_{0.70}$	$82.47_{1.01}$	$84.25_{1.22}$	$74.72_{0.67}$	$75.21_{1.42}$	$75.90_{0.33}$	$78.54_{0.43}$	$87.09_{0.98}$	$86.63_{1.08}$	$80.48_{1.32}$	$84.21_{1.20}$	$84.32_{0.45}$
PerFedAvg	$78.43_{0.43}$	$80.51_{1.26}$	$83.44_{0.47}$	$68.42_{1.74}$	$70.34_{0.48}$	$73.21_{1.11}$	$74.53_{1.04}$	$77.38_{0.22}$	$81.77_{0.48}$	$81.58_{0.52}$	$85.91_{0.30}$	$87.22_{0.38}$
FedALA	$81.01_{1.50}$	$83.63_{0.43}$	$85.99_{1.04}$	$74.99_{0.42}$	$76.87_{1.26}$	$79.45_{1.65}$	$81.98_{0.44}$	$88.25_{2.21}$	$92.31_{0.84}$	$83.42_{0.04}$	$87.24_{1.96}$	$87.99_{1.48}$
FedCLIP	$82.22_{0.89}$	$84.58_{1.26}$	$85.72_{1.25}$	$74.45_{1.21}$	$75.33_{0.51}$	$77.54_{0.34}$	$81.37_{0.35}$	$89.46_{0.12}$	$92.86_{0.99}$	$84.47_{1.29}$	$89.36_{1.38}$	$89.55_{0.95}$
FACMIC	$84.21_{1.03}$	$85.41_{0.33}$	$87.27_{0.19}$	$75.03_{1.57}$	$76.47_{1.98}$	$79.94_{1.20}$	$82.11_{0.37}$	$89.12_{0.77}$	$92.78_{0.57}$	$85.28_{0.38}$	$89.74_{1.42}$	$93.21_{1.20}$
FedTCA (Ours)	$\mathbf{90.21}_{0.22}$	$\mathbf{90.79}_{0.18}$	$\mathbf{94.33}_{1.20}$	$\mathbf{83.99}_{0.21}$	$\mathbf{88.02}_{0.41}$	$\mathbf{90.26}_{0.13}$	$\mathbf{92.42}_{0.26}$	$\mathbf{95.77}_{0.09}$	$\mathbf{98.47}_{0.37}$	$\mathbf{95.67}_{0.27}$	$\mathbf{97.67}_{0.23}$	$\mathbf{97.99}_{0.21}$

Compare with Prompt Learning-based FL. We compared with Prompt Learning-based FL for CLIP adaptation: PromptFL [13], PromptFL+FT [13], pFedPrompt [12], and FedTPG [22], which replace original prompts with learnable vectors to facilitate adaptation. **Table 2** demonstrates that our FedTCA, significantly surpasses these benchmarks. The superiority of FedTCA stems from employing LoRA in the CLIP encoder, which accelerates cross-domain generalization from natural to medical contexts at minimal cost (1.1% of total parameters). Additionally, our generation of hard textual medical prompts, restyled for learnable context-aware flexibility, offers more domain-specific insights compared to methods that merely substitute textual prompts with learnable parameters.

Table 2. Results on Prompt Learning-based FL.

Dataset	BloodMNIST			TissueMNIST		
Method	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
PromptFL	$80.01_{0.29}$	$81.24_{1.126}$	$83.15_{1.00}$	$73.97_{0.49}$	$75.01_{0.39}$	$75.26_{0.34}$
PromptFL+FT	$83.42_{0.88}$	$84.90_{0.14}$	$86.20_{1.03}$	$74.73_{0.27}$	$76.11_{0.29}$	$78.99_{0.20}$
pFedPrompt	$86.88_{0.34}$	$86.94_{0.31}$	$89.87_{0.54}$	$78.56_{0.45}$	$80.88_{0.46}$	$83.25_{0.12}$
FedTPG	$87.05_{0.87}$	$87.79_{0.32}$	$90.12_{0.23}$	$77.97_{0.51}$	$81.04_{0.68}$	$83.27_{0.45}$
FedTCA (Ours)	$\mathbf{90.21}_{0.22}$	$\mathbf{90.79}_{0.18}$	$\mathbf{94.33}_{1.20}$	$\mathbf{83.99}_{0.21}$	$\mathbf{88.02}_{0.41}$	$\mathbf{90.26}_{0.13}$
Dataset	OrganMNIST (2D)			OrganMNIST (3D)		
Method	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$
PromptFL	$80.46_{0.21}$	$83.27_{1.53}$	$87.86_{0.71}$	$82.46_{0.63}$	$86.54_{0.89}$	$87.21_{1.04}$
PromptFL+FT	$82.86_{1.12}$	$86.00_{0.31}$	$88.24_{1.73}$	$84.53_{1.12}$	$88.62_{0.42}$	$89.99_{0.04}$
pFedPrompt	$87.26_{0.12}$	$90.02_{0.42}$	$94.80_{0.55}$	$86.64_{0.65}$	$90.21_{2.03}$	$92.63_{1.42}$
FedTPG	$88.04_{0.24}$	$90.34_{0.54}$	$94.96_{0.37}$	$87.00_{0.23}$	$90.07_{0.55}$	$92.79_{0.82}$
FedTCA (Ours)	$\mathbf{92.42}_{0.26}$	$\mathbf{95.77}_{0.09}$	$\mathbf{98.47}_{0.37}$	$\mathbf{95.67}_{0.27}$	$\mathbf{97.67}_{0.23}$	$\mathbf{97.99}_{0.21}$

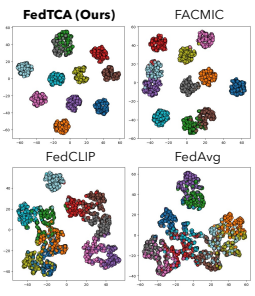

Fig. 3. t-SNE visualization [19] on OrganMNIST (2D) - $\lambda = 0.5$.

Ablation Studies. Table 3 demonstrates that our method outperforms original similarity and classic OT alignment. This stems from addressing the limitations of non-optimal transport, which tends to average the distances between feature maps and cues, thereby underscoring the importance of optimal transport in ensuring robust visual alignment. Additionally, the consistent superiority of unbalanced over classical OT across all scenarios supports our method's effectiveness. In addition, we evaluate the performance under different action scope of LoRA in **Table 4**. Unlike using LoRA on individual encoders or standard fine-tuning, our strategy achieves superior cross-domain adaptation without substantially increasing costs, offering an optimal performance-efficiency trade-off.

Table 3. Ablation on Alignment. Results based on BloodMNIST with $\lambda = 0.5$.

Ablation Case	Alignment	Blood	Tissue
FedTCA-A	Original Similarity	$90.33_{0.20}$	$85.24_{0.31}$
FedTCA-B	Classical OT	$91.26_{0.42}$	$87.93_{0.16}$
FedTCA	Our TCA	$\mathbf{94.33}_{1.20}$	$\mathbf{90.26}_{0.13}$

Table 4. Ablation on LoRA. Results based on BloodMNIST with $\lambda = 0.5$.

Ablation Case	$\boldsymbol{E}_{\text{image}}$	$\boldsymbol{E}_{\text{text}}$	Acc.	Train Params#
Standard FT	✘	✘	$90.04_{0.77}$	0.13 M
LoRA-B	✔	✘	$92.36_{0.40}$	0.79 M
LoRA-C	✘	✔	$91.93_{0.16}$	0.79 M
Original	✔	✔	$\mathbf{94.33}_{1.20}$	1.46 M

Hyperparamerter Sensitivity. Tables 5, 6 and 7 assess the impact of three critical hyperparameters in FedTCA. Experiments were conducted with $\lambda = 0.5$, and the gray shading indicates our default setup. Key findings include: (1) Higher r improves performance but also increase computational costs; the default setting of $r = 4$ provides the optimal performance-cost trade-off. (2) Increasing the regularization weight η leads to reduced performance as higher η enforce stricter visual-text semantic alignment, thus restricting the advantages of a more diverse distribution. (3) global and local prompt transport plans collaboratively align vision representations according to OT constraints. However, decreasing γ focuses alignment on narrower regions of the object, thereby reducing performance.

Table 5. Impact of rank r in LoRA. TP#: Trainable Param.

r	Blood	Tissue	TP# (M)
4	$94.33_{1.20}$	$90.26_{0.13}$	1.46
8	$94.65_{0.17}$	$90.49_{0.25}$	2.78
32	$94.99_{0.41}$	$90.52_{0.44}$	10.75

Table 6. Impact of different η in Eq. 3.

η	Blood	Tissue
0.1	$94.33_{1.20}$	$90.26_{0.13}$
0.3	$94.01_{0.67}$	$89.24_{0.30}$
0.5	$93.88_{0.28}$	$90.01_{0.54}$

Table 7. Impact of different γ in our TCA.

γ	Blood	Tissue
1.0	$\mathbf{94.33}_{1.20}$	$90.26_{0.13}$
0.9	$94.25_{0.10}$	$88.56_{0.09}$
0.7	$92.73_{0.65}$	$89.99_{0.26}$

Scaling Discussions Real-world medical FL is constrained by limited client data [4, 6, 3, 24], hindering performance gains. FedTCA's lightweight communication and LoRA enable scaling to boost performance without added overhead.

4 Conclusion

This paper introduces FedTCA, a novel framework that enhances pretrained VLMs through a Prompt Restyling strategy, replacing standard prompts with learnable, domain-specific versions and incorporating a low-rank adaptation for improved cross-domain knowledge transfer. Additionally, FedTCA features the Twin Cross-modal Alignment (TCA), which addresses learning biases from non-IID data by optimizing the balance between global and personalized knowledge through conceptualizing knowledge transfer as an optimal transport problem, enhancing local visual-text alignment. Evaluated extensively on real-world medical imaging datasets, FedTCA has demonstrated superior effectiveness.

Acknowledgments. Thanks to the reviewers for their constructive feedback. This work is supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515011438, 2024A1515510031), the National Natural Science Foundation of China (No. 42105145), and the Scientific Foundation for Youth Scholars of Shenzhen University (No. 868-000001033384).

Disclosure of Interests. The authors declare no competing interests.

References

1. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)
2. Chen, H., Zhao, B., Yue, G., Liu, W., Lv, C., Wang, R., Zhou, F.: Clip-medfake: Synthetic data augmentation with ai-generated content for improved medical image classification. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 3854–3860. IEEE (2024)
3. Chen, S., Long, G., Jiang, J., Zhang, C.: Personalized adapter for large meteorology model on devices: Towards weather foundation models. arXiv preprint arXiv:2405.20348 (2024)
4. Chen, S., Long, G., Shen, T., Jiang, J., Zhang, C.: Federated prompt learning for weather foundation models on devices. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 5772–5780 (2024)
5. Chen, S., Ren, S., Wang, G., Huang, M., Xue, C.: Interpretable cnn-multilevel attention transformer for rapid recognition of pneumonia from chest x-ray images. IEEE Journal of Biomedical and Health Informatics **28**(2), 753–764 (2023)
6. Chen, S., Shu, T., Zhao, H., Wang, J., Ren, S., Yang, L.: Free lunch for federated remote sensing target fine-grained classification: A parameter-efficient framework. Knowledge-Based Systems **294**, 111694 (2024)
7. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
8. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Dykstra, R.L.: An algorithm for restricted least squares regression. Journal of the American Statistical Association **78**(384), 837–842 (1983)

10. Fallah, A., Mokhtari, A., Ozdaglar, A.: Personalized federated learning: A meta-learning approach. arXiv preprint arXiv:2002.07948 (2020)
11. GDPR, G.D.P.R.: General data protection regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (2016)
12. 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)
13. Guo, T., Guo, S., Wang, J., Tang, X., Xu, W.: Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. IEEE Transactions on Mobile Computing **23**(5), 5179–5194 (2023)
14. 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 preprint arXiv:2106.09685 (2021)
15. Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K.: Are natural domain foundation models useful for medical image classification? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7634–7643 (2024)
16. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021)
17. Lin, L., Liu, Y., Wu, J., Cheng, P., Cai, Z., Wong, K.K., Tang, X.: Fedlppa: Learning personalized prompt and aggregation for federated weakly-supervised medical image segmentation. arXiv preprint arXiv:2402.17502 (2024)
18. Lu, W., Hu, X., Wang, J., Xie, X.: Fedclip: Fast generalization and personalization for clip in federated learning. arXiv preprint arXiv:2302.13485 (2023)
19. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
20. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
21. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning **11**(5-6), 355–607 (2019)
22. Qiu, C., Li, X., Mummadi, C.K., Ganesh, M.R., Li, Z., Peng, L., Lin, W.Y.: Text-driven prompt generation for vision-language models in federated learning. arXiv preprint arXiv:2310.06123 (2023)
23. 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)
24. Ren, S., Hu, Y., Chen, S., Wang, G.: Federated distillation for medical image classification: Towards trustworthy computer-aided diagnosis. arXiv preprint arXiv:2407.02261 (2024)
25. Sun, G., Mendieta, M., Luo, J., Wu, S., Chen, C.: Fedperfix: Towards partial model personalization of vision transformers in federated learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4988–4998 (2023)

26. T Dinh, C., Tran, N., Nguyen, J.: Personalized federated learning with moreau envelopes. Advances in neural information processing systems **33**, 21394–21405 (2020)
27. Wu, Y., Desrosiers, C., Chaddad, A.: Facmic: Federated adaptative clip model for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
28. 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)
29. Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., Guan, H.: Fedala: Adaptive local aggregation for personalized federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11237–11244 (2023)
30. Zhang, J., Hua, Y., Wang, H., Song, T., Xue, Z., Ma, R., Guan, H.: Fedcp: Separating feature information for personalized federated learning via conditional policy. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3249–3261 (2023)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
32. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)