

Global and Local Prompts Cooperation via Optimal Transport for Federated Learning

Hongxia Li¹ Wei Huang² Jingya Wang¹ Ye Shi^{1,*}

¹ShanghaiTech University, Shanghai, China

²RIKEN Center for Advanced Intelligence Project, Japan

{lihx2,wangjingya,shiye}@shanghaitech.edu.cn, wei.huang.vr@riken.jp

<https://github.com/HongxiaLee/FedOTP>

Abstract

Prompt learning in pretrained visual-language models has shown remarkable flexibility across various downstream tasks. Leveraging its inherent lightweight nature, recent research attempted to integrate the powerful pretrained models into federated learning frameworks to simultaneously reduce communication costs and promote local training on insufficient data. Despite these efforts, current federated prompt learning methods lack specialized designs to systematically address severe data heterogeneities, e.g., data distribution with both label and feature shifts involved. To address this challenge, we present Federated Prompts Cooperation via Optimal Transport (FedOTP), which introduces efficient collaborative prompt learning strategies to capture diverse category traits on a per-client basis. Specifically, for each client, we learn a global prompt to extract consensus knowledge among clients, and a local prompt to capture client-specific category characteristics. Unbalanced Optimal Transport is then employed to align local visual features with these prompts, striking a balance between global consensus and local personalization. Extensive experiments on datasets with various types of heterogeneities have demonstrated that our FedOTP outperforms the state-of-the-art methods.

1. Introduction

Federated learning [48] is a distributed machine learning framework that enables decentralized collaboration among participants without sharing their training data. However, current federated learning methods involve high training and communication costs due to the need to update and share model parameters with the server. This constraint has typically restricted these methods to modest backbone architectures, hindering their feature capacity and resulting in

performance limitations and training instability [69].

Recently, vision-language pre-trained models like Contrastive Language-Image Pretraining (CLIP) [55] have shown potential in learning robust and versatile representations suitable for various image distributions, aligning with the objectives of federated learning. However, the substantial communication overhead between the server and clients renders training CLIP in a federated learning environment. Additionally, overfitting concerns may arise when large-scale models are trained with limited client data. Prompt learning [42, 72, 73] provides a flexible way to adapt pre-trained models to downstream tasks by training only additional parameters. This enables prompts to capture task-specific information while guiding the fixed model’s performance. Leveraging its lightweight nature, prior research [26, 71] has explored the integration of prompt learning into federated learning frameworks to overcome the problems outlined above.

In real-world scenarios, client data often exhibits variations in domain discrepancies (feature shift) [40] or imbalanced class distributions (label shift) [36]. Simply applying the FedAvg [48] method on prompts [26] across all clients tends to deviate from their local distribution, leading to unsatisfactory performance. Hence, it’s crucial to develop specialized personalized federated prompt learning approaches to effectively address data heterogeneity. pFedPrompt [25] introduced personalization into federated prompt learning by maintaining personalized attention modules to generate spatial visual features locally while learning user consensus through shared text prompts. However, in the presence of a significant label shift or notable feature shift, merely learning a shared prompt in the language modality is inadequate.

To resolve these limitations, we propose simultaneously learning both a shared global prompt and a personalized local prompt for each client in the local training phase. After local training, the local prompt is retained locally, while the global prompt is transmitted to the server to aggregate

*Corresponding author.

with prompts from other clients using FedAvg [48]. In this manner, the client owns the capacity to acquire consensus knowledge among clients from the global prompt, while also being able to discern client-specific user traits through the local prompt.

To further achieve a balance between global consensus and local personalization, we introduce Federated Prompts Cooperation via Optimal Transport (FedOTP). FedOTP utilizes Optimal Transport (OT) [31] to align local visual features with both global and local textual features through an adaptive transport plan, promoting fine-grained matching across vision and language modalities and strengthening collaboration between the global and local prompts. The adaptive OT transport plan can provide resilience to visual misalignment and effective adaptation to feature shifts. It’s worth noting that the standard OT formulation imposes two hard equality constraints on the transport plan, leading to each image patch being assigned to prompts. This may potentially cause prompts to capture some class-irrelevant information from the image and consequently influence the final results. To mitigate this, we consider employing unbalanced OT by relaxing one of the equality constraints, allowing prompts to focus solely on the most relevant image patches rather than the entire content of the image. For an efficient solution, we apply a fast implementation of Dijkstra’s algorithm [15] in our FedOTP, enabling swift convergence and focusing on the core area of the image during iterations.

Our main contributions are summarized as follows:

- We are the first to explore the mechanism of prompts’ cooperation in federated learning where severe data heterogeneity is present. More precisely, we train a global prompt to learn consensus information among clients, and a local prompt to capture client-specific category characteristics at the same time.
- We propose FedOTP, a federated learning framework utilizing unbalanced OT to enhance the cooperation between global and local prompts. By aligning the local visual features and these two textual prompts, our FedOTP can effectively deal with severe data heterogeneity.
- We conducted extensive experiments on widely adopted datasets in various data heterogeneity with domain discrepancy and imbalanced class distribution, and the significant result improvement verifies the superiority of our FedOTP. In addition, we demonstrated the ability of FedOTP to balance consensus and local personalization through visualizations.

2. Related Work

2.1. Personalized Federated Learning

Personalized federated learning (PFL) is a highly regarded research field because of its potential to address statisti-

cal and systemic heterogeneity across clients. Various approaches have been proposed in prior research to achieve PFL. The most common method involves the inclusion of regularization terms in the loss function [38, 39, 61], and fine-tuning the global model on clients’ local datasets [16, 33, 46, 67]. Additionally, some works focus on explicitly seeking a trade-off between the global model and the local models [7, 27, 41, 47]. To enhance adaptability to diverse data distributions, certain techniques have delved into clustering methods for client grouping [30, 56, 65]. Leveraging the relationships and data distribution among clients, methods like FedPAC [68], and FedDisco [70] introduce novel weighted aggregation techniques to promote intensive collaboration among similar clients. Furthermore, some researchers have explored the decomposition of model parameters into base layers and personalized layers. For instance, FedPer [1], FedRep [11], and FedBABU [51] learn personalized classifier heads locally while sharing the base layers, and FedTP [35] learns personalized self-attention layers for each client.

The methods mentioned above primarily target label shift data heterogeneity. However, they may not perform well when substantial domain differences exist among clients. In dealing with these feature shifts, FedBN [40] employs local batch normalization to mitigate the feature shift before model averaging, while PartialFed [60] extends this strategy by selecting personalized parameters according to distinct feature traits of different clients. Besides, FedPCL [62] enhances each client’s ability to utilize pre-trained models by extracting client-specific and class-relevant information. Our FedOTP explores the cooperation between global and local prompts to effectively address both label shift and feature shift data heterogeneity.

2.2. Prompt-based Federated Learning

Prompt learning, originating from NLP models, has expanded to Vision Language Models. Initial methods like CLIP [55] involved manually crafted templates, while recent approaches concentrate on learning prompts in a continuous embedding space. CoOp [73] fine-tunes CLIP with continuous prompt vectors. Based on this, plenty of studies [6, 29, 32, 43, 45, 72] have been introduced to enhance the effectiveness of prompt learning. To accelerate the global aggregation and handle situations with insufficient user data, FedPrompt [71] and PromptFL [26] have introduced prompt learning into Federated Learning. Based on these two works, several methods have made substantial progress in various domains. For instance, FedPR [21] focuses on learning federated visual prompts within the null space of the global prompt for MRI reconstruction. Based on CLIP, FedAPT [59] introduces a federated adaptive prompt tuning algorithm for cross-domain federated image classification, and FedCLIP [44] utilizes an attention-based

adapter to optimize the utilization of pre-trained model information. To tackle statistical heterogeneity among clients, pFedprompt [25] maintains a non-parametric personalized attention module for each client to generate locally personalized spatial visual features, and pFedPG [69] designs a client-specific prompt generator at the server to create personalized prompts. While these works show the potential of prompt learning in Federated Learning, there remains a deficiency in technical enhancements tailored to PFL scenarios. Compared with these methods, our FedOTP employs OT to balance the global consensus and local personalization from the collaboration of global and local prompts.

2.3. Optimal Transport

Initially developed as a solution to optimize the cost of moving multiple items concurrently, Optimal Transport (OT) [31] has gained significant attention in the machine learning and computer vision community. To accelerate the convergence and efficiently deal with large-scale problems, [12] introduced Sinkhorn’s algorithm [58] for computing an approximate transport coupling with entropic regularization. Classical OT lacks the flexibility for partial displacement in the transport plan. Unbalanced OT addresses this by relaxing the equality constraint, and by incorporating soft penalties based on Kullback-Leibler divergence [22], it can be efficiently solved through the generalized Sinkhorn’s algorithm [9]. Given its remarkable ability in distribution matching, OT has been applied in various theoretical and practical tasks, including domain adaptation [4, 13, 18], learning with noisy labels [5, 20], causal discovery [37, 63, 64], federated learning [8, 17] and so on. In prompt learning field, PLOT [6] proposes to learn multiple prompt sets for diverse contextual representations and use OT to align the features of vision and language modalities. Different from PLOT, we employ unbalanced OT to enhance the cooperation between global and local prompts by relaxing one of the equality constraints, which allows prompts to concentrate exclusively on the most relevant image patches.

3. Preliminaries

3.1. Prompt Learning

To adapt pre-trained models like CLIP [55] to downstream tasks, prompt learning methods [72, 73] provide an efficient way by training a few parameters in the prompt. Within the CLIP model, the textual prompts are manually crafted using class labels $y \in \{1, 2, \dots, K\}$ (e.g., “a photo of a $\langle \text{classname} \rangle$ ”) representing K classes. These textual prompts are then tokenized and projected into word embeddings $W = \{w_1, w_2, \dots, w_L\} \in \mathbb{R}^{L \times d_l}$ where L is the number of word embeddings and d_l denotes its dimension. To learn the context prompts, we interpose $s(\leq L)$ learn-

able vectors $\{p_i \in \mathbb{R}^{d_l}\}_{i=1}^s$ in the language branch. Consequently, the textual prompt can be formulated as $P_k = \{w_1, p_1, \dots, p_s, w_{s+2}, \dots, w_L\} \in \mathbb{R}^{L \times d_l}$, where we use $\{p_1, \dots, p_s\}$ in place of $\{w_2, \dots, w_{s+1}\}$ to be consistent with previous works. Denote the fixed text encoder as $h(\cdot)$ and image encoder as $g(\cdot)$, and the prediction probabilities for each category are computed with the input prompt P_k of class k and image x through matching scores:

$$q(y = k|\mathbf{x}) = \frac{\exp(\text{sim}(g(x), h(P_k))/\tau)}{\sum_{c=1}^K \exp(\text{sim}(g(x), h(P_c))/\tau)}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ denotes a metric function such as cosine similarity, and τ represents the temperature of Softmax. Next, we optimize the learnable parameters $\{p_i\}_{i=1}^s$ by minimizing the cross-entropy loss

$$\ell_{CE} = -\frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{k=1}^K y_{\mathbf{x},k} q(y = k|\mathbf{x}), \quad (2)$$

where $y_{\mathbf{x}}$ is a one-hot label vector.

3.2. Optimal Transport

Optimal Transport is a constrained optimization problem aiming to efficiently transfer probability mass between two distributions. Here we briefly recall its formulation of the discrete situation. Given two probability simplex vectors α and β and a cost matrix $C \in \mathbb{R}^{|\alpha| \times |\beta|}$, OT aims to find the optimal transport plan T by minimizing the following objective:

$$d_C(\alpha, \beta) = \min_{T \in U(\alpha, \beta)} \langle C, T \rangle, \quad (3)$$

$$U(\alpha, \beta) = \left\{ T \in \mathbb{R}_+^{|\alpha| \times |\beta|} \mid T \mathbf{1}_{|\beta|} = \alpha, T^\top \mathbf{1}_{|\alpha|} = \beta \right\},$$

where $\langle \cdot, \cdot \rangle$ is Frobenius dot-product, $U(\alpha, \beta)$ denotes the solution space of T , and $\mathbf{1}_d$ is a d -dimensional vector of ones. Directly optimizing the OT problem would be time-consuming. Sinkhorn algorithm [12] introduces an entropic regularization term for fast optimization. The regularized OT formulation can be expressed as: $\min_{T \in U(\alpha, \beta)} \langle C, T \rangle + \lambda \langle T, \log T \rangle$, where $\lambda \geq 0$ is a hyper-parameter. In light of this, the optimal transport plan T^* has been shown to be unique with the form $T^* = \text{diag}(u^{(\tilde{t})}) \exp(-C/\lambda) \text{diag}(v^{(\tilde{t})})$, where \tilde{t} represents the iteration and in each iteration $u^{(\tilde{t})} = u/(\exp(-C/\lambda)v^{(\tilde{t}-1)})$ and $v^{(\tilde{t})} = v/(\exp(-C/\lambda)^\top u^{(\tilde{t})})$.

4. Methodology

In this section, we present the design of our FedOTP framework, illustrated in Figure 1. To achieve a balance between global consensus and local personalization, FedOTP utilizes unbalanced Optimal Transport to strengthen the collaboration between global and local prompts, effectively addressing both label shift and feature shift data heterogeneity.

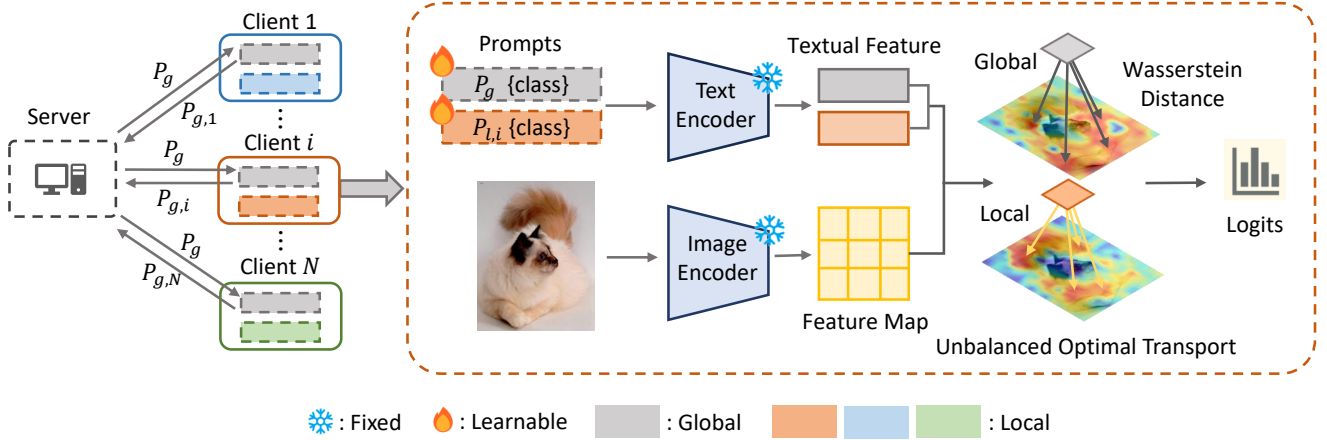


Figure 1. Overview of our FedOTP. On the left, clients transmit global prompts to the server for aggregation while retaining local prompts locally. The right shows the workflow of Global-Local prompt cooperation mechanism, which employs unbalanced Optimal Transport to align visual feature maps with each prompt.

4.1. Federated Learning with Global and Local Prompts

Consider a federated learning scenario involving N clients and a central server, and each client i holds a local dataset $D_i = \{(x_i^j, y_i^j)\}_{j=1}^{m_i}$ ($i = 1, \dots, N$) containing m_i samples. Let $D = \{D_1, D_2, \dots, D_N\}$ represent the total datasets where each dataset is derived from a distinct data distribution \mathcal{D}_i , and let C_t represent the set of selected clients participating at communication round t .

Aiming to reduce communication costs and address data heterogeneity, each client is equipped with a pre-trained CLIP model and a prompt learner in our federated learning setup. Through the prompt learner, every client learns both a shared global prompt and a personalized local prompt, allowing clients to extract more individualized insights while maintaining a degree of consensus among them. Specifically, for each client i , the prompt P_i comprises a global prompt P_g and a personalized prompt $P_{l,i}$, denoted as $P_i = [P_g, P_{l,i}]$. During each communication round t , client i initializes the prompt with $P_i^{t,0} = [P_g^{t-1}, P_{l,i}^{t-1}]$. Then the global and local prompts are jointly updated through gradient descent $P_i^{t,r} = P_i^{t,r-1} - \eta \nabla \mathcal{L}_{\mathcal{D}_i}(P_i^{t,r-1})$ for R iterations locally. After local training, only the updated global prompt $P_{g,i}^{t,R}$ is transmitted to the server for aggregation to learn global consensus among clients, while the personalized prompt is retained locally to capture client-specific category characteristics. The process of aggregation can be expressed as:

$$P_g^t = \sum_{i \in C_t} \frac{m_i}{\sum_{j \in C_t} m_j} P_{g,i}^{t,R}. \quad (4)$$

Then the objective function of our FedOTP can be for-

mulated as:

$$\min_{P_g, \{P_{l,i}\}_{i=1}^N} \sum_{i=1}^N \frac{m_i}{\sum_{j \in C_t} m_j} \mathcal{L}_{\mathcal{D}_i}(P_g, P_{l,i}), \quad (5)$$

with $\mathcal{L}_{\mathcal{D}_i}(P_g, P_{l,i}) = \mathbb{E}_{(x_i^j, y_i^j) \in \mathcal{D}_i} \ell(f(P_g, P_{l,i}; x_i^j), y_i^j)$, where $f(P_g, P_{l,i}; \cdot)$ represents the personalized model for client i , and $\ell(\cdot, \cdot)$ denotes the cross-entropy loss function as introduced in Eq. (2).

4.2. Global-Local Prompt Cooperation by Unbalanced Optimal Transport

In this subsection, we introduce the details of the prompt learning process for each client, which leverages unbalanced OT to integrate insights learned from both global and local prompts. To be specific, as shown in Figure 1, we initialize prompts P_g and $P_{l,i}$ as $\{w_1, p_1, \dots, p_s, \dots, w_L\}$ where w_i represents the word embedding and p_i signifies learnable vectors. With the text encoder $h(\cdot)$, we obtain a global textual feature $H_{k,g} = h(P_{g,k}) \in \mathbb{R}^{d_f}$ and a local textual feature $H_{k,l} = h(P_{l,i,k}) \in \mathbb{R}^{d_f}$ for each class k , and the combination of these two features is represented as $H_k = [H_{k,g}, H_{k,l}]$ for convenience. Uploading an image x to the image encoder $g(\cdot)$, we derive a set of visual features $G = g(x) \in \mathbb{R}^{(V+1) \times d_f}$, which consists of a class token $G_c \in \mathbb{R}^{d_f}$ and a feature map $G_m \in \mathbb{R}^{V \times d_f}$.

We consider learning an optimal transport plan T that aligns both global and local textual features H_k with visual feature map G_m . By representing features as samples from discrete distributions, the cost matrix can be represented by the cosine distance between H_k and G_m as $C = 1 - G_m^T H_k \in \mathbb{R}^{V \times 2}$, then the optimization objective

of unbalanced optimal transport is formulated as:

$$d_{C,k}(\alpha, \beta) = \min_{T \in U(\alpha, \beta)} \langle C, T \rangle, \quad (6)$$

$$U(\alpha, \beta) = \{T \in \mathbb{R}_+^{V \times 2} \mid T \mathbb{1}_2 \leq \alpha, T^\top \mathbb{1}_V = \beta\},$$

where $\alpha \in \mathbb{R}^V$ and $\beta \in \mathbb{R}^2$ are essentially marginal probability vectors which satisfy $\|\alpha\|_1 \geq \|\beta\|_1 = \gamma$ ($\gamma \in [0, 1]$). The difference between Eq. (6) and formulation in PLOT [6] lies in their use of classical OT with two hard equality constraints as Eq. (3). This forces prompts to map to each image patch, potentially causing them to capture some class-irrelevant information from the image and thereby influencing the final results. In contrast, our FedOTP 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 , FedOTP owns the ability to regulate the mapping size of prompts on the feature map.

For fast optimization, we add an entropic regularization term to Eq. (6), and the objective function is formulated as follows:

$$d_{C,k}(\alpha, \beta) = \min_{T \in U(\alpha, \beta)} \langle C, T \rangle + \lambda \langle T, \log T \rangle. \quad (7)$$

In line with [2], we can further reformulate Eq. (7) as a Kullback-Leibler (KL) projection, and the solution space $U(\alpha, \beta)$ is then defined as the intersection of two convex but not affine sets:

$$d_{C,k}(\alpha, \beta) = \min_{T \in U(\alpha, \beta)} \lambda \text{KL}(T \mid e^{-C/\lambda}), \quad (8)$$

$$\mathcal{C}_1 \triangleq \{T \in \mathbb{R}_+^{V \times 2} \mid T \mathbb{1}_2 \leq \alpha\},$$

$$\mathcal{C}_2 \triangleq \{T \in \mathbb{R}_+^{V \times 2} \mid T^\top \mathbb{1}_V = \beta\}.$$

To solve Eq. (8), we employ a rapid implementation of Dijkstra’s algorithm [15] as introduced in [5], which efficient scales iterative KL projection between \mathcal{C}_1 and \mathcal{C}_2 by leveraging matrix-vector multiplications exclusively. Initializing $Q = \exp(-C/\lambda)$ and $v^{(0)} = \mathbb{1}_2$, a fast optimization solution is achieved within a few iterations as:

$$T^* = \text{diag}(u^{(\tilde{t})}) Q \text{diag}(v^{(\tilde{t})}), \quad (9)$$

where \tilde{t} is the iteration, and in each iteration $u^{(\tilde{t})} = \min(\mathbb{1}_V / Q_\alpha v^{(\tilde{t}-1)}, \mathbb{1}_V)$ and $v^{(\tilde{t})} = \mathbb{1}_2 / Q_\beta^\top u^{(\tilde{t})}$ with $Q_\alpha = Q / \text{diag}(\alpha) \mathbb{1}_{V \times 2}$ and $Q_\beta^\top = Q^\top / \text{diag}(\beta) \mathbb{1}_{V \times 2}$. The details of this algorithm are shown in Appendix Section A.1.

By Eq. (9), we obtain the optimal transport plan T^* and the final Wasserstein distance $d_{C,k}$, then the matching scores in Eq. (1) is replaced by the following prediction probability:

$$q(y = k \mid \mathbf{x}) = \frac{\exp((1 - d_{C,k})/\tau)}{\sum_{c=1}^K \exp((1 - d_{C,c})/\tau)}. \quad (10)$$

After obtaining $q(y = k \mid \mathbf{x})$, we fix the transport plan T^* and optimize learnable vectors $\{p_a\}_{a=1}^s$ in both global and local prompts simultaneously for client i through cross entropy as described in Eq. (2). Then the global prompt $P_{g,i}$ is sent to the server for aggregation utilizing Eq. (4) with the local prompt retained locally. During local training via OT, the final prediction probability of FedOTP is a synthesis of information derived from both the global and the local prompts. This avoids a straightforward addition of the outcomes from the two prompts, fostering a comprehensive and collaborative learning process. Due to page limitation, the algorithm box is deferred to the Appendix Section A.2.

4.3. Generalization Bound

We analyze the generalization bound of our FedOTP in this section. Before starting the analysis, we first introduce some assumptions as follows.

Assumption 1 (Lipschitz Conditions) Let $\mathcal{D}_1, \dots, \mathcal{D}_N$ denote the real data distribution of each client and $\mathcal{L}_{\mathcal{D}_i}(P_g, P_{l,i}) = \mathbb{E}_{(x_i^j, y_i^j) \in \mathcal{D}_i} \ell(f(P_g, P_{l,i}; x_i^j), y_i^j)$ be the expected loss. We assume the following Lipschitz conditions hold:

$$|\ell(f((P; x), y) - \ell(f((P'; x), y))| \leq L \|f((P; x), y) - f((P'; x), y)\|, \quad (11a)$$

$$\|f(P_g, P_{l,i}) - f(P'_g, P_{l,i})\| \leq L_g \|P_g - P'_g\|, \quad (11b)$$

$$\|f(P_g, P_{l,i}) - f(P_g, P'_{l,i})\| \leq L_{l,i} \|P_{l,i} - P'_{l,i}\|. \quad (11c)$$

Assumption 2 Since the convergence of global prompt has been proved in [26], we assume $\|\hat{P}_g - P_g^*\|_2 \leq A_g$ for convenience. And we assume local prompts $P_{l,i}$ are bounded in a ball of radius $A_{l,i}$, which can be denoted as $\|\hat{P}_{l,i} - P_{l,i}^*\|_2 \leq A_{l,i}$.

Leveraging above assumptions, we can derive the following theorem:

Theorem 1 (Generalization Bound of FedOTP)

Suppose $\hat{\mathcal{D}}_1, \dots, \hat{\mathcal{D}}_N$ denote empirical data distribution of N clients with learned parameters \hat{P}_g and $\hat{P}_{l,i}$, and P_g^* and $P_{l,i}^*$ are optimal parameters for the real distribution $\mathcal{D}_1, \dots, \mathcal{D}_N$. Let \mathcal{H} represent the personalized hypothesis and d denote the VC-dimension of \mathcal{H} . Suppose all the clients participate at every communication round and Assumptions 1 and 2 hold, with probability at least $1 - \delta$, we have

$$\left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| \leq \sqrt{\frac{M}{2} \log \frac{N}{\delta}} + \sqrt{\frac{dN}{M} \log \frac{eM}{d}} + L(L_g A_g + L_l A_l), \quad (12)$$

Table 1. The results of our FedOTP and the benchmark methods on the Pathological Non-IID setting with non-overlapping over 10 clients.

Methods	Food101	DTD	Caltech101	Flowers102	OxfordPets
Local Training					
Zero-Shot CLIP [55]	75.27±0.05	40.21±0.12	85.14±0.24	62.17±0.12	84.47±0.10
CoOp [73]	82.54±2.42	82.69±0.63	90.41±0.44	88.23±0.76	94.52±1.30
Prompt-based Federated Learning					
PromptFL [26]	74.81±0.64	50.46±0.54	87.90±0.54	73.68±1.58	88.17±1.18
PromptFL+FT [23]	77.16±1.56	53.74±1.36	89.70±0.25	72.31±0.91	91.23±0.50
PromptFL+FedProx [38]	73.96±0.75	50.89±0.71	87.80±1.10	74.14±0.65	87.25±1.48
PromptFL+FedPer [1]	71.29±1.87	50.23±0.82	86.72±1.45	72.11±1.35	89.50±1.62
PromptFL+FedAMP [30]	74.48±1.71	47.16±0.92	87.31±1.60	69.10±0.13	80.21±0.44
pFedPrompt [25]	92.26±1.34	77.14±0.09	96.54±1.31	86.46±0.15	91.84±0.41
FedOTP (Ours)	92.73±0.15	87.67±0.70	97.02±0.36	96.23±0.44	98.82±0.11

where $M = \sum_{i=1}^N m_i$, and we denote $L_l = \sqrt{\sum_{i=1}^N L_{l,i}^2}$ and $A_l = \sqrt{\sum_{i=1}^N A_{l,i}^2}$ for simplicity. Theorem 1 indicates that the performance of FedOTP trained on the empirical distribution relates to the model complexity and Lipschitz assumptions. More details and proof of Theorem 1 are provided in the Appendix Section E.

5. Experiments

In this section, we conduct comprehensive experiments to numerically evaluate our FedOTP in the scenarios of heterogeneous data distribution.

5.1. Experimental Setup

Datasets and Data Heterogeneity. We evaluated the performance of our framework on nine public benchmark datasets with different types of heterogeneity, including label shift and feature shift. To investigate label shift heterogeneity, we selected two types of datasets. Following previous research [25, 26], we utilized five representative visual classification datasets to simulate datasets with limited samples: Food101 [3], DTD [10], Caltech101 [19], Flowers102 [50], and OxfordPets [52]. Referring to these datasets as the CLIP dataset for convenience, we utilized a Pathological setting by randomly allocating a distinct number of non-overlapping classes to each client. We also employed two image benchmark datasets: CIFAR-10, and CIFAR-100 [34]. We considered the Dirichlet Distribution as introduced in [35, 57] where the datasets are partitioned randomly among clients using a symmetric Dirichlet distribution with $\alpha = 0.3$. For feature shift heterogeneity, we evaluated our framework on the following two datasets with multiple domains: DomainNet [54] with 6 domains, and Office-Caltech10 [24] with 4 domains. In line with prior studies [40, 62], each client participating in the federated learning

system is assigned data from one of these distinct domains. A detailed introduction of each dataset and details about the Non-IID settings can be found in the Appendix Section B.1.

Baselines. We compared FedOTP with three kinds of baselines: (1) Local training methods: (i) Zero-shot CLIP [55] with hand-crafted text prompt templates; (ii) CoOp [73] with learnable prompt vectors trained on each client locally. (2) Existing prompt-based federated learning: (i) PromptFL [26] using a unified prompt learned across clients using FedAvg [48]; (ii) pFedPrompt [25] learning a shared prompt with personalized visual attention modules for each client. (3) Four adapted baseline methods derived from traditional PFL techniques, including PromptFL+FT [23], PromptFL+FedProx [38], PromptFL+FedPer [1] and PromptFL+FedAMP [30], as introduced in [25].

Implementation Details. To simulate federated learning in various scenarios, we consider the following two settings: (1) $n = 10$ clients with a full $r = 100\%$ partition, (2) $n = 100$ clients with a $r = 10\%$ partition. We employ SGD optimizer with a learning rate $lr = 0.001$ and local epoch $R = 5$ for CLIP datasets while $R = 1$ for other cases. The communication round is set to $T = 10$ for CLIP datasets with 10 clients and $T = 150$ for CIFAR-10/CIFAR-100 datasets with 100 clients. We present the results using two representative backbones, ResNet50 [28] and ViT_B16 [14], defaulting to ViT_B16 if not explicitly specified. All experiments are conducted with Pytorch [53] on NVIDIA A40 GPUs. More implementation details can be found in the Appendix Section B.2.

5.2. Performance Evaluation

Evaluation Protocol. We evaluated the models on each client’s private test data whose distribution is consistent with its training set. The reported results are the average test accuracy across all clients from three different seeds.

Table 2. Experimental results on DomainNet dataset with feature & label shifts.

Datasets	DomainNet						
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
Local Training							
Zero-Shot CLIP [55]	8.72±1.73	12.48±3.78	8.53±4.32	9.31±0.69	9.13±2.55	11.96±2.80	10.02±2.65
CoOp [73]	44.40±14.89	45.68±16.53	47.21±18.20	41.13±20.62	48.02±24.49	39.47±5.68	44.32±16.74
Prompt-based Federated Learning							
PromptFL [26]	9.31±6.53	12.58±9.91	8.23± 8.47	14.79±12.07	9.37±10.82	7.48±11.32	10.29±10.35
PromptFL+FedProx [38]	9.84±6.60	11.16±11.17	10.64±6.79	13.40±16.09	9.39±7.69	6.78±11.76	10.20±10.99
FedOTP (Ours)	46.14±6.53	60.14±18.23	45.2±16.86	38.66±7.60	49.30±17.80	49.02±24.22	48.08±15.21

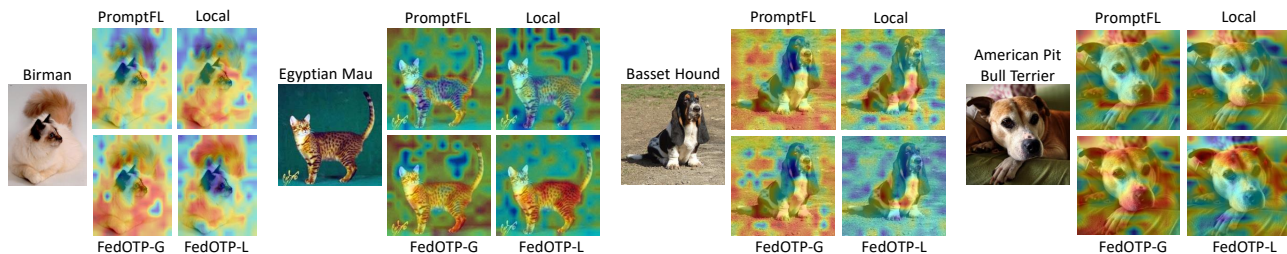


Figure 2. Heatmaps of similarity between text features and image feature maps for different methods on 4 categories in OxfordPets dataset. “FedOTP-G” denotes the results from the global prompt and “FedOTP-L” refers to the local prompt.

Table 3. The results of our FedOTP and the benchmark methods on Dirichlet settings in CIFAR-10 and CIFAR-100 over 100 clients.

Methods	CIFAR-10	CIFAR-100
Local Training		
Zero-Shot CLIP [55]	87.71±0.68	64.92±0.53
CoOp [73]	93.11±0.39	74.83±0.45
Prompt-based Federated Learning		
PromptFL [26]	92.30±0.87	73.67±0.56
PromptFL+FedProx [38]	91.83±0.47	71.11±0.91
FedOTP (Ours)	96.05±0.12	78.03±0.08

Model Evaluation on Label Shifts. We first measured the performance of FedOTP against baselines on datasets with label shifts. The experimental results on CLIP datasets and CIFAR-10/CIFAR-100 datasets are summarized in Table 1 and Table 3. For easy comparison, Table 1 reports results utilizing ResNet50 as the backbone, maintaining consistency with [25]. As shown in Table 1, our FedOTP outperforms state-of-the-art algorithms by a large margin across all datasets, which confirms the effectiveness of our Global-Local prompt cooperation mechanism to handle label shift scenarios. Remarkably, while both PromptPer (which splits the learnable prompt vector into “base+personalized” vectors) and pFedPrompt (utilizing a shared prompt with a personalized attention module in the vision modal) experience significant declines when datasets are altered, Fe-

dOTP exhibits slight fluctuations. This verifies the robustness of our method across diverse scenarios. Table 3 shows the results of our FedOTP and benchmark methods on CIFAR-10/CIFAR-100 datasets under Dirichlet setting over 100 clients with 10% partition. Even in this scenario with Dirichlet settings and a large number of clients, FedOTP consistently outperforms the baseline methods, further highlighting the superiority of our approach.

Model Evaluation on Feature & Label Shifts. In this set of experiments, we explored scenarios involving both feature shifts and label shifts by partitioning data within a domain into five clients based on the Dirichlet distribution with $\alpha = 0.1$. We analyzed the mean and variance of clients in the same domain, and the outcomes for the DomainNet dataset are summarized in Table 2. In the presence of two types of data heterogeneity, our method performs favorably against baselines. We observe that, with significant data heterogeneity across clients, traditional federated learning methods experience a pronounced performance decline compared to local training. In contrast, our FedOTP exhibits superior performance, achieving a 3.7% increase in average accuracy on each domain. Additional experimental results on feature shifts and in the Office-Caltech10 dataset are available in the Appendix Section C.1 and C.2.

Visualization. We first investigated the interplay between global and local prompts by representing the similarity between text features and image feature maps as a heatmap on OxfordPets [52] dataset. To be specific, we compared the

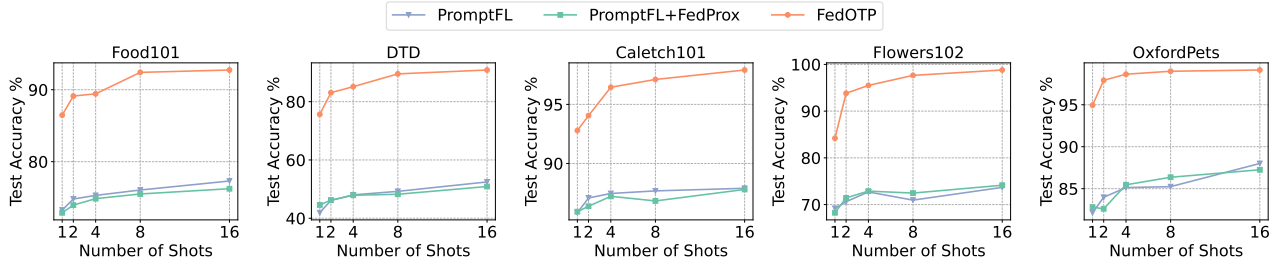


Figure 3. Performance with the different number of shots.

Table 4. Quantitative comparisons on the Pathological Non-IID setting across different numbers of shots over 10 clients.

Datasets	Food101		DTD		Caltech101		Flowers102		OxfordPets	
	2	8	2	8	2	8	2	8	2	8
FedOTP (Similarity Averaging)	83.38±0.54	87.59±1.05	81.01±0.23	88.17±0.73	92.68±0.44	96.73±0.29	91.73±0.68	97.09±0.18	96.23±0.25	98.34±0.15
FedOTP (Classical OT)	88.07±0.63	89.77±0.62	81.42±0.99	88.43±0.45	93.17±0.68	96.80±0.23	92.84±1.34	97.07±0.25	96.55±0.26	98.51±0.27
FedOTP (Unbalanced OT)	89.12±0.28	92.94±0.18	85.50±0.35	90.25±0.74	95.05±0.49	97.34±0.18	93.96±0.48	98.23±0.32	97.73±0.57	99.02±0.38

heatmaps of our FedOTP with PromptFL and Local Training using CoOp, and the original images and corresponding heatmaps are illustrated in Figure 2. We observed that global prompts of FedOTP might concentrate more on common features, like limbs and facial characteristics, while local prompts tended to capture client-specific details such as the special tail of “Birman”, unique patterns of “Egyptian Mau” and “Basset Hound”, and the distinct dent on the head of “Terrier”. This demonstrates the effectiveness of FedOTP in balancing global consensus and local personalization. More visualization results on transport plans of our FedOTP will be given in the Appendix Section D.1.

5.3. Ablation Study

Impact of Number of Shots. Following the few-shot evaluation setting adopted in [25, 26], we further investigated the impact of the number of shots in FedOTP. To analyze this, we varied the number of shots during the training process within the range of [1, 2, 4, 8, 16]. Results are summarized in Figure 3, where the horizontal axis denotes the number of shots and the vertical axis represents the average test accuracy. We observe that as the number of shots increases, the corresponding performance of each method gradually improves. However, our FedOTP consistently exhibits a dominant edge over methods with a shared global prompt in all scenarios.

Effectiveness of the Unbalanced OT. In this subsection, we explored the effectiveness of OT on two variants of FedOTP briefly described below: (1) FedOTP (Similarity Averaging): removing OT in FedOTP and matching global and local prompts with visual feature maps by averaging similarities of each visual-textual pair; (2) FedOTP (Classical OT): employing classical OT during the matching process. The results in Table 4 demonstrate the effectiveness

of utilizing OT to align feature maps with global and local prompts compared to FedOTP (Similarity Averaging) in almost all cases, particularly on the Food101 dataset. This is because the absence of OT leads to the feature map’s distance from prompts reverting to the mean distance of each feature-prompt pair, highlighting the crucial role of OT in providing resilience to visual misalignment. In addition, the persistent superiority of unbalanced OT over classical OT across all scenarios serves as a compelling testament to the effectiveness of our approach.

6. Conclusion

In this paper, we proposed Federated Prompts Cooperation via Optimal Transport (FedOTP), a novel framework designed to facilitate efficient model personalization across heterogeneous clients. In our approach, each client is equipped with both a global prompt and a local prompt, and then unbalanced Optimal Transport is utilized to align local visual features with these prompts, fostering enhanced collaboration between global and local prompts. With fine-grained matching facilitated by OT, FedOTP effectively addresses data heterogeneity characterized by domain discrepancy and imbalanced class distributions. Our extensive experiments across diverse datasets consistently demonstrate the superior performance of FedOTP in tackling both label shifts and feature shifts, which verifies the effectiveness of our Global-Local prompt cooperation mechanism via OT. Through visualization results, we confirmed that global prompts learned by FedOTP concentrated on common features among all clients, while local prompts captured individual client-specific details. In future work, we aim to investigate the generalization capabilities of our method on novel clients unseen during the training process.

Acknowledgement

This work was supported by NSFC (No.62303319), Shanghai Sailing Program (22YF1428800, 21YF1429400), Shanghai Local College Capacity Building Program (23010503100), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

- [1] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. **2, 6**
- [2] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. **5**
- [3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. **6, 2**
- [4] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022. **3**
- [5] Wanxing Chang, Ye Shi, and Jingya Wang. CSOT: Curriculum and Structure-Aware Optimal Transport for Learning with Noisy Labels. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. **3, 5, 1**
- [6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. **2, 3, 5**
- [7] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021. **2**
- [8] Yi-Han Chiang, Koudai Terai, Tsung-Wei Chiang, Hai Lin, Yusheng Ji, and John CS Lui. Optimal Transport based One-Shot Federated Learning for Artificial Intelligence of Things. *IEEE Internet of Things Journal*, 2023. **3**
- [9] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. **3**
- [10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. **6, 2**
- [11] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021. **2**
- [12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. **3, 1**
- [13] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 447–463, 2018. **3**
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. **6**
- [15] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983. **2, 5**
- [16] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020. **2**
- [17] Farzan Farnia, Amirhossein Reisizadeh, Ramtin Pedarsani, and Ali Jadbabaie. An optimal transport approach to personalized federated learning. *IEEE Journal on Selected Areas in Information Theory*, 3(2):162–171, 2022. **3**
- [18] Kilian Fatras, Thibault SÉjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021. **3**
- [19] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. **6, 2**
- [20] Chuanwen Feng, Yilong Ren, and Xike Xie. OT-Filter: An Optimal Transport Filter for Learning With Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16164–16174, 2023. **3**
- [21] Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning Federated Visual Prompt in Null Space for MRI Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8064–8073, 2023. **2**
- [22] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. *Advances in neural information processing systems*, 28, 2015. **3**
- [23] Karan N. Chadha Gary Cheng and John C. Duchi. Fine-tuning is Fine in Federated Learning. *CoRR*, abs/2108.07313, 2021. **6**
- [24] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012. **6, 2**

- [25] Tao Guo, Song Guo, and Junxiao Wang. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, 2023. 1, 3, 6, 7, 8
- [26] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. PromptFL: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023. 1, 2, 5, 6, 7, 8, 4
- [27] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020. 2
- [28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [29] Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, et al. Hyperprompt: Prompt-based task-conditioning of transformers. In *International Conference on Machine Learning*, pages 8678–8690. PMLR, 2022. 2
- [30] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized Cross-Silo Federated Learning on Non-IID Data. In *AAAI*, pages 7865–7873, 2021. 2, 6
- [31] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006. 2, 3
- [32] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2
- [33] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 2
- [35] Hongxia Li, Zhongyi Cai, Jingya Wang, Jiangnan Tang, Weiping Ding, Chin-Teng Lin, and Ye Shi. FedTP: Federated Learning by Transformer Personalization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 6
- [36] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 1
- [37] Qian Li, Zhichao Wang, Shaowu Liu, Gang Li, and Guandong Xu. Causal optimal transport for treatment effect estimation. *IEEE transactions on neural networks and learning systems*, 2021. 3
- [38] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. 2, 6, 7, 4
- [39] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021. 2
- [40] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations*, 2020. 1, 2, 6
- [41] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 2
- [42] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023. 1
- [43] Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, and Chengguang Gui. Hierarchical Prompt Learning for Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10888–10898, 2023. 2
- [44] Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. FedCLIP: Fast Generalization and Personalization for CLIP in Federated Learning. *arXiv preprint arXiv:2302.13485*, 2023. 2
- [45] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 2
- [46] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020. 2, 10
- [47] Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized Federated Learning through Local Memorization. In *International Conference on Machine Learning*, pages 15070–15092. PMLR, 2022. 2
- [48] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 6
- [49] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. 8
- [50] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6, 2
- [51] Jaehoon Oh, Sangmook Kim, and Se-Young Yun. FedBABU: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021. 2
- [52] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6, 7, 2

- [53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [54] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6, 2
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7, 4
- [56] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020. 2
- [57] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR, 2021. 6
- [58] Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967. 3
- [59] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Cross-domain federated adaptive prompt tuning for clip. *arXiv preprint arXiv:2211.07864*, 2022. 2
- [60] Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. Partialfed: Cross-domain personalized federated learning via partial initialization. *Advances in Neural Information Processing Systems*, 34:23309–23320, 2021. 2
- [61] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020. 2
- [62] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in Neural Information Processing Systems*, 35:19332–19344, 2022. 2, 6
- [63] William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021. 3
- [64] Ruibo Tu, Kun Zhang, Hedvig Kjellström, and Cheng Zhang. Optimal transport for causal discovery. *arXiv preprint arXiv:2201.09366*, 2022. 3
- [65] Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10043–10052, 2023. 2
- [66] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6
- [67] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019. 2
- [68] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023. 2
- [69] Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19159–19168, 2023. 1, 3
- [70] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. FedDisco: Federated Learning with Discrepancy-Aware Collaboration. *arXiv preprint arXiv:2305.19229*, 2023. 2
- [71] Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Reduce communication costs and preserve privacy: Prompt tuning method in federated learning. *arXiv preprint arXiv:2208.12268*, 2022. 1, 2
- [72] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 3
- [73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 6, 7, 4

Global and Local Prompts Cooperation via Optimal Transport for Federated Learning Supplementary Materials

Supplementary organization:

A Method Details	1
A.1 Efficient Scaling Dykstra’s Algorithm	1
A.2 Training Process	2
B Experimental Details	2
B.1. Details of Dataset Setup	2
B.2 Implementation Details	3
C Additional Experiments Results	4
C.1. Model Evaluation on Feature Shifts	4
C.2. Model Evaluation on Feature & Label Shifts	4
C.3. Effect of Parameter γ in Unbalanced OT	4
C.4. Effect of Heterogeneity in Label Distribution	5
C.5. Learning Curves	5
D Visualization	5
D.1. Visualizations of Transport Plans	5
D.2 T-SNE Projection of Prompts.	6
E Generalization Bound	8
E.1. Key Lemmas	8
E.2. Proof of Theorem 1	9

A. Method Details

A.1. Efficient Scaling Dykstra’s Algorithm

As introduced in [5], Problem (8) can be solved by a fast implementation of Dykstra’s Algorithm by only performing matrix-vector multiplications, which is very similar to the widely-used and efficient Sinkhorn Algorithm [12]. The details of this algorithm are shown in Algorithm 1.

Algorithm 1: Efficient Scaling Dykstra’s algorithm

Input: Cost matrix C , marginal constraints vectors α and β , entropic regularization weight λ .

- 1 Initialize: $Q \leftarrow e^{-C/\lambda}$, $v^{(0)} \leftarrow \mathbb{1}_\beta$, $\Delta_v = \infty$, $\epsilon = 0.001$;
 - 2 Compute: $Q_\alpha \leftarrow \frac{Q}{\text{diag}(\alpha)\mathbb{1}_{|\alpha|\times|\beta|}}$, $Q_\beta^\top \leftarrow \frac{Q^\top}{\text{diag}(\beta)\mathbb{1}_{|\beta|\times|\alpha|}}$;
 - 3 **for** $n = 1, 2, 3, \dots$ **do**
 - 4 $u^{(n)} \leftarrow \min \left(\frac{\mathbb{1}_{|\alpha|}}{Q_\alpha v^{(n-1)}}, \mathbb{1}_{|\alpha|} \right)$;
 - 5 $v^{(n)} \leftarrow \frac{\mathbb{1}_{|\beta|}}{Q_\beta^\top u^{(n)}}$;
 - 6 $\Delta_v = |v^{(n)} - v^{(n-1)}|$;
 - 7 **if** $\Delta_v < \epsilon$ **then**
 - 8 | break
 - 9 **end**
 - 10 **end**
 - 11 **return** $\text{diag}(u^{(n)})Q\text{diag}(v^{(n)})$
-

A.2. Training Process

Here, we provide detailed descriptions of the algorithm for our FedOTP, as shown in Algorithms 2. For each communication round t , the selected clients perform local training by training global and local prompts $P_i^t = [P_g^t, P_{l,i}^t]$ through unbalanced OT at the same time. Then the updated global prompts $P_{g,i}^t$ are sent to the server for aggregation.

Algorithm 2: FedOTP: Federated Prompts Cooperation via Optimal Transport

Input: Communication rounds T , local epochs R , client number N , local dataset D_i , sample numbers m_i , pre-trained CLIP model $g(\cdot)$ and $h(\cdot)$, class number K , learning rate η , temperature of Softmax τ .

- 1 Initialize parameters $P_i^0 = [P_g^0, P_{l,i}^0]$;
- 2 **for** each communication round $t \in \{1, \dots, T\}$ **do**
- 3 Sample a client set $C^t \subset \{1, \dots, N\}$;
- 4 **for** each client $i \in C^t$ **do**
- 5 Initialize $P_i^{t,0} = [P_g^{t-1}, P_{l,i}^{t-1}]$;
- 6 **for** each local epoch $r \in \{1, \dots, R\}$ **do**
- 7 Sample a mini-batch $B_i \in D_i$;
- 8 Obtain a visual feature map G_m with the visual encoder $g(x)(x \in B_i)$;
- 9 Obtain textual features H_k of each class with the textual encoder $\{h(P_{i,k}^{t,r-1})\}_{k=1}^K$;
- 10 Calculate the cost matrix $C_k = 1 - G_m^\top H_k$ of each class;
- 11 Solve Problem (8) through Algorithm 1 and obtain Wasserstein distance $d_{C,k} = \langle T_k^*, C_k \rangle$;
- 12 Calculate the classification probability $q(y = k | \mathbf{x}) = \frac{\exp((1-d_{C,k})/\tau)}{\sum_{c=1}^K \exp((1-d_{C,c})/\tau)}$;
- 13 Update the parameters of prompts $P_i^{t,r} \leftarrow P_i^{t,r-1} - \eta \nabla \mathcal{L}_{\mathcal{D}_i}(P_i^{t,r-1})$;
- 14 **end**
- 15 **end**
- 16 Aggregate the global prompt $P_g^t = \sum_{i \in C^t} \frac{m_i}{\sum_{j \in C^t} m_j} P_{g,i}^{t,R}$;
- 17 **end**
- 18 **return** $P_i = [P_g, P_{l,i}]$

B. Experimental Details

B.1. Details of Dataset Setup

We select nine representative visual classification datasets as our benchmark. The detailed statistics of each dataset are shown in Table A1, including the original tasks, the number of classes, the size of training and testing samples, and the number of domains. As for datasets with multiple domains, Office-Caltech10 is a standard benchmark dataset consisting of four

Table A1. The detailed statistics of datasets used in experiments.

Dataset	Task	Classes	Training Size	Testing Size	Domains
Caltech101 [19]	Object recognition	100	4,128	2,465	1
Flowers102 [50]	Fine-grained flowers recognition	102	4,093	2,463	1
OxfordPets [52]	Fine-grained pets recognition	37	2,944	3,669	1
Food101 [3]	Fine-grained food recognition	101	50,500	30,300	1
DTD [10]	Texture recognition	47	2,820	1,692	1
CIFAR10 [34]	Image Classification	10	50,000	10,000	1
CIFAR100 [34]	Image Classification	100	50,000	10,000	1
DomainNet [54]	Image recognition	10	18278	4573	6
Office-Caltech10 [24]	Image recognition	10	2025	508	4

domains, namely Amazon, Caltech, DSLR, and WebCam, which are acquired using different camera devices or in different real environments with various backgrounds. DomainNet is a large-scale dataset consisting of six domains, namely Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. We selected 10 classes from each of these two datasets for training. Some examples of raw instances of these two datasets can be found in Figure A1. For a clearer illustration, we visualize the three Non-IID settings employed in our paper in Figure A2.

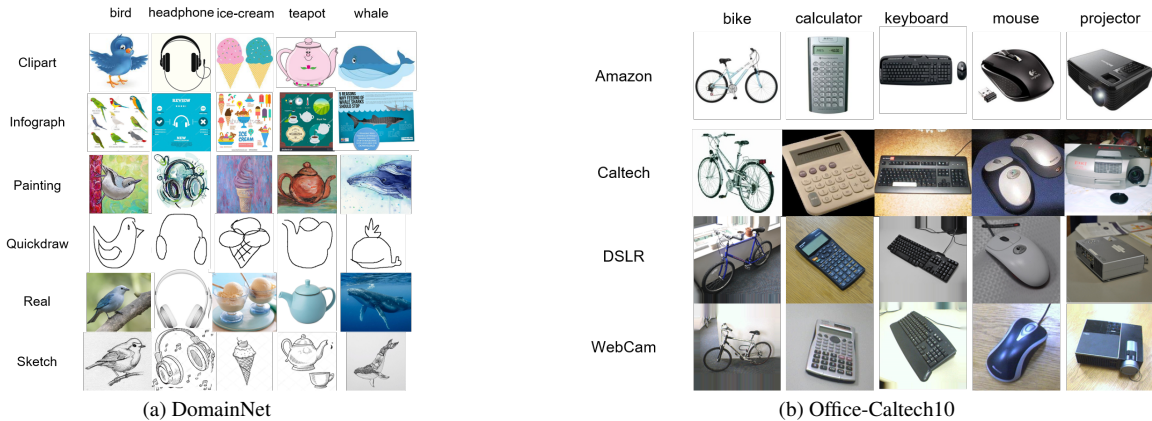


Figure A1. Examples of raw instances from two datasets with multiple domains: DomainNet (left) and Office-Caltech10 (right). We present five classes for each dataset to show the feature shift across their sub-datasets.



Figure A2. Visualization of three Non-IID settings on the Office-Caltech10 dataset. Each dot represents a set of samples within specific classes assigned to a client, with the dot size indicating the number of samples. The feature shifts are denoted by different colors.

B.2. Implementation Details

All input images across datasets are resized to 224×224 pixels and further divided into 14×14 patches with a dimension of 768. Regarding the hyperparameters for solving OT, we set the entropic regularization weight in Problem Eq. (8) as $\lambda = 0.1$ for all datasets. The maximum iteration number n for Algorithm 1 is set to 100, and we implement early stopping when the absolute update value Δ_v is less than 0.001. For the setting of learnable prompts, the length of prompt vectors s is set to 16 with a dimension of 512, “end” token position, and “random” initialization. Batch sizes are set to 32 for training and 100 for testing.

C. Additional Experiments Results

C.1. Model Evaluation on Feature Shifts

In Table A2, we compared the performance on Office-Caltech10 and DomainNet datasets under the presence of feature shift, where each client is assigned data from distinct domains while sharing the same label distribution. Our method achieved the highest average accuracies 99.16% and 94.55% on Office-Caltech10 and DomainNet, respectively.

Table A2. Experimental results on Office-Caltech10 and DomainNet datasets with feature shift.

Datasets Domains	Office-Caltech10					DomainNet						
	A	C	D	W	Avg.	C	I	P	Q	R	S	Avg.
<i>Local Training</i>												
Zero-Shot CLIP [55]	19.3	18.2	21.9	18.6	19.50	49.92	47.15	53.63	31.3	48.4	50.18	46.76
CoOp [73]	96.38	97.24	100	98.31	97.98	98.32	83.01	98.18	82.37	98.21	97.70	92.95
<i>Prompt-based Federated Learning</i>												
PromptFL [26]	96.41	96.39	96.90	100	97.42	98.23	79.91	97.89	66.52	96.83	97.31	89.45
PromptFL+FedProx [38]	97.93	97.21	96.89	100	98.01	98.45	72.32	96.00	63.51	96.08	98.04	87.40
FedOTP (Ours)	97.92	98.68	100	100	99.16	98.93	84.52	98.89	87.87	98.64	98.42	94.55

C.2. Model Evaluation on Feature & Label Shifts

In this set of experiments, we investigated scenarios involving both feature shifts and label shifts by dividing data within a domain into three clients based on the Dirichlet distribution with $\alpha = 0.1$ for the Office-Caltech10 dataset. We calculated the mean and standard deviation of clients in the same domain, and the outcomes are presented in Table A3. Comparing these results with those in Table A2, we can observe that the introduction of label shift leads to a performance decrease across all methods, with federated learning methods employing a shared prompt experiencing the most significant decline. In spite of this, our FedOTP consistently achieves the highest average accuracy, demonstrating its capability to utilize both global and local prompts to capture general domain-invariant and specific domain-specific knowledge for effective adaptation to extreme data heterogeneity.

Table A3. Experimental results on Office-Caltech10 dataset with feature & label shifts.

Datasets Domains	Office-Caltech10				
	Amazon	Caltech	DSLR	Webcam	Avg.
<i>Local Training</i>					
Zero-Shot CLIP [55]	8.45±1.49	6.01±4.25	12.92±9.15	6.48±4.82	8.46±6.26
CoOp [73]	25.59±6.60	36.23±16.97	30.30±5.18	22.56±5.46	28.67±11.13
<i>Prompt-based Federated Learning</i>					
PromptFL [26]	10.92±4.36	10.37±12.63	15.45±15.34	15.90±17.33	13.16±13.60
PromptFL+FedProx [38]	11.05±4.70	12.04±10.65	19.70±21.74	12.56±12.72	13.84±14.29
FedOTP (Ours)	23.59±4.74	31.64±5.25	43.94±5.67	35.51±9.19	33.67±9.76

C.3. Effect of Parameter γ in Unbalanced OT

In this subsection, we delved into the effect of parameter γ in unbalanced OT, which regulates the mapping size of prompts on the feature map. We conducted experiments on the Pathological Non-IID setting across four datasets with varying numbers of shots and different values of the parameter γ in our FedOTP. Specifically, we set $R = 5$ and $T = 10$ for these experiments. The results presented in Table A4 reveal a notable trend: as the parameter γ decreases, the overall performance initially increases and subsequently decreases. Interestingly, the majority of optimal results are observed at $\gamma = 0.8$ or $\gamma = 0.7$. This observation implies that the optimal alignment between global and local prompts and the feature map is achieved when the mapping size of prompts on the feature map is around 70% – 80%. Consequently, we adopt $\gamma = 0.8$ in other experiments.

Table A4. Quantitative comparisons on the Pathological Non-IID setting across varying numbers of shots with different parameter γ in our FedOTP over 10 clients.

Dataset	shot number	1	0.9	0.8	0.7	0.6	0.5
DTD	1 shot	74.22±0.75	73.72±0.79	75.75±0.64	72.81±0.42	77.36±0.98	77.22±1.46
	2 shots	81.89±0.76	84.03±0.57	84.64±0.29	85.50±0.35	80.39±0.24	82.47±0.40
	4 shots	85.06±0.91	85.75±0.63	86.69±0.61	87.67±0.70	86.58±0.51	85.86±0.44
	8 shots	88.64±0.31	88.22±0.30	89.77±0.24	90.25±0.74	89.17±0.53	89.67±0.51
	16 shots	90.51±0.11	91.02±0.48	91.31±0.59	90.94±0.25	89.97±0.34	90.33±0.51
Caltech101	1 shot	89.68±1.19	92.13±0.58	92.54±0.71	91.53±0.55	90.10±1.74	90.67±0.44
	2 shots	95.05±0.49	93.89±0.35	94.45±0.32	94.37±0.43	93.89±0.65	94.68±0.92
	4 shots	96.02±0.36	96.64±0.41	97.02±0.36	96.68±0.46	96.38±0.42	96.66±0.37
	8 shots	96.74±0.21	96.79±0.24	96.91±0.16	96.95±0.26	97.22±0.33	97.34±0.18
	16 shots	97.72±0.14	97.69±0.17	97.39±0.11	97.58±0.23	97.74±0.19	97.83±0.18
Flowers102	1 shot	86.68±1.93	85.77±0.74	87.42±0.92	88.43±0.90	89.14±1.18	85.56±1.21
	2 shots	93.09±1.26	93.96±0.48	93.31±0.55	93.13±0.26	93.70±0.49	93.56±0.86
	4 shots	95.46±0.55	96.23±0.44	95.51±0.30	95.89±0.50	96.17±0.47	96.16±0.34
	8 shots	97.53±0.24	97.49±0.19	98.23±0.32	98.11±0.27	97.24±0.28	97.40±0.64
	16 shots	98.86±0.15	98.30±0.55	99.11±0.11	98.88±0.17	99.03±0.12	98.93±0.18
OxfordPets	1 shot	95.82±1.16	94.26±0.38	96.18±0.71	96.37±0.79	94.21±0.53	95.97±0.42
	2 shots	97.73±0.57	96.12±0.32	97.50±1.02	97.49±0.45	97.60±0.40	97.27±0.19
	4 shots	98.11±1.15	98.46±0.64	98.82±0.11	98.51±0.10	98.52±0.25	98.43±0.28
	8 shots	98.73±0.27	99.02±0.38	98.71±0.16	98.74±0.18	98.54±0.22	98.63±0.13
	16 shots	99.04±0.16	98.82±0.25	99.27±0.23	99.21±0.19	99.04±0.27	98.81±0.21

C.4. Effect of Heterogeneity in Label Distribution

In addressing the core challenge of data heterogeneity in personalized federated learning, FedOTP consistently outperforms benchmark methods across various settings. Now, we investigated the effect of heterogeneity in label distribution by considering a range of α values of Dirichlet distribution, specifically $\alpha \in \{0.1, 0.3, 0.5, 1, 5, 10\}$ for CIFAR-100 datasets. It's worth noting that a smaller α implies a higher degree of data heterogeneity in these experiments. The results presented in Table A5 clearly indicate that as the degree of data heterogeneity increases, the performance of federated learning methods with a shared prompt decreases while the performance of CoOp and our FedOTP improves. Among these methods, FedOTP outperforms them in every case and demonstrates remarkable robustness. These findings underscore the effectiveness of FedOTP in overcoming label distribution heterogeneity across a diverse range of scenarios.

C.5. Learning Curves

To assess the convergence of our method, we plotted test accuracy curves with $R = 1$ and $T = 50$ for different methods across four datasets, as illustrated in Figure A3. Compared to other methods, FedOTP exhibits notable characteristics of accelerated convergence and enhanced stability, evident from the smaller fluctuations in test accuracy.

D. Visualization

D.1. Visualizations of Transport Plans

To facilitate a comparative analysis between FedOTP, PromptFL, and Local models, we examined visualizations of similarity between textual features and feature maps in Figure 2. Here, we provided visualization examples showcasing transport plans T associated with global and local prompts of FedOTP across different γ values. We converted each transport plan into colorful heatmaps, resized and overlaid them on the original image. The comparisons between heatmaps of transport plans

Table A5. Quantitative comparisons on CIFAR-100 dataset with different α of the Dirichlet setting.

Dataset # α	CIFAR-100					
	0.1	0.3	0.5	1	5	10
Local Training						
Zero-Shot CLIP [55]	65.22±0.32	64.92±0.53	65.78±0.41	63.93±0.16	64.01±0.27	65.07±0.35
CoOp [73]	62.01±0.29	74.83±0.45	51.72±0.42	47.03±0.37	41.03±0.23	41.37±0.19
Prompt-based Federated Learning						
PromptFL [26]	72.45±0.64	73.67±0.56	74.37±0.18	73.95±0.14	74.68±0.05	74.43±0.08
PromptFL+FedProx [38]	72.57±0.54	71.11±0.91	74.45±0.19	74.19±0.06	74.23±0.09	74.53±0.07
FedOTP (Similarity Averaging)	78.68±0.17	75.70±0.27	75.28±0.12	74.88±0.16	74.48±0.05	74.31±0.39
FedOTP (Classical OT)	79.93±0.19	77.86±0.09	75.76±0.12	75.38±0.08	75.01±0.05	74.73±0.05
FedOTP (Unbalanced OT)	80.56±0.12	78.03±0.08	76.75±0.10	76.17±0.13	75.75±0.03	75.52±0.06

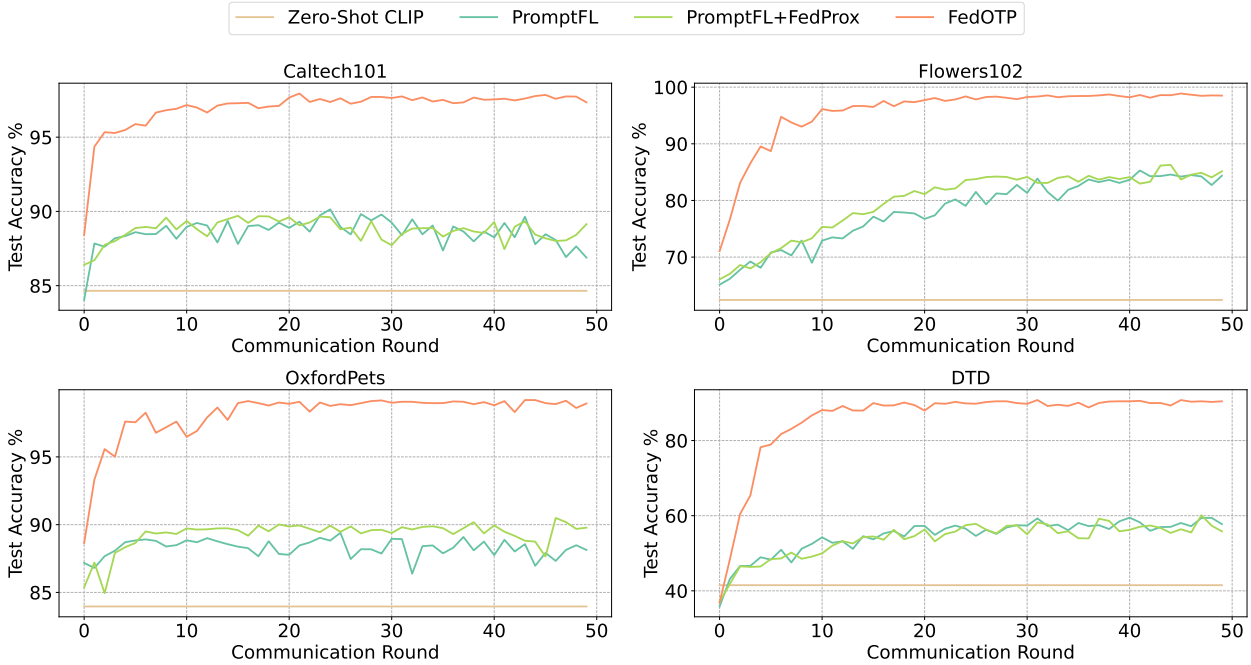


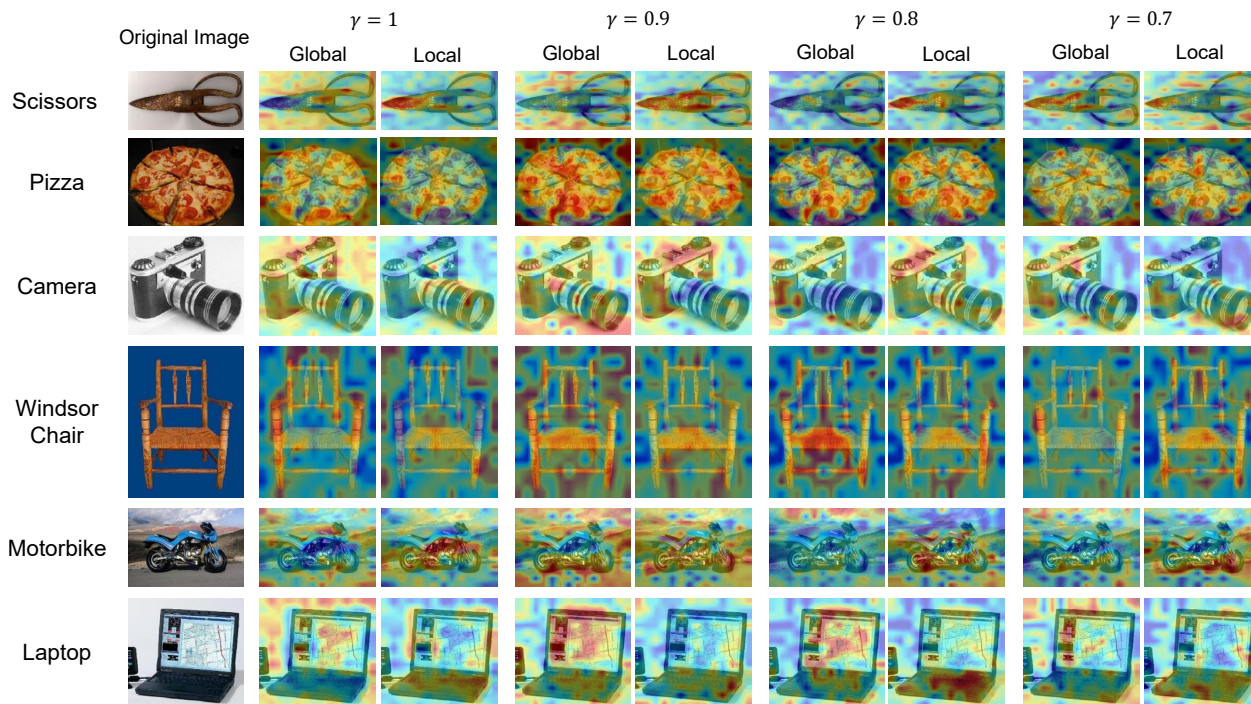
Figure A3. Accuracy curves and convergence behavior of FedOTP and other baselines on four datasets over 10 clients.

with different γ in Caltech101 dataset is presented in Figure A4. Upon observation, we noted that when $\gamma = 1$, the transport plans related to global and local prompts exhibit complementary, as each image patch is assigned to prompts due to the equality constraints of OT. This may result in the integration of objects and backgrounds on the global part, observable in classes like “Camera” and “Laptop”. In contrast, as γ decreases, prompts focus on a smaller range of patches and primarily center on the patches of main objects rather than backgrounds, further supporting the claim that FedOTP can effectively regulate the mapping size of prompts on the feature map.

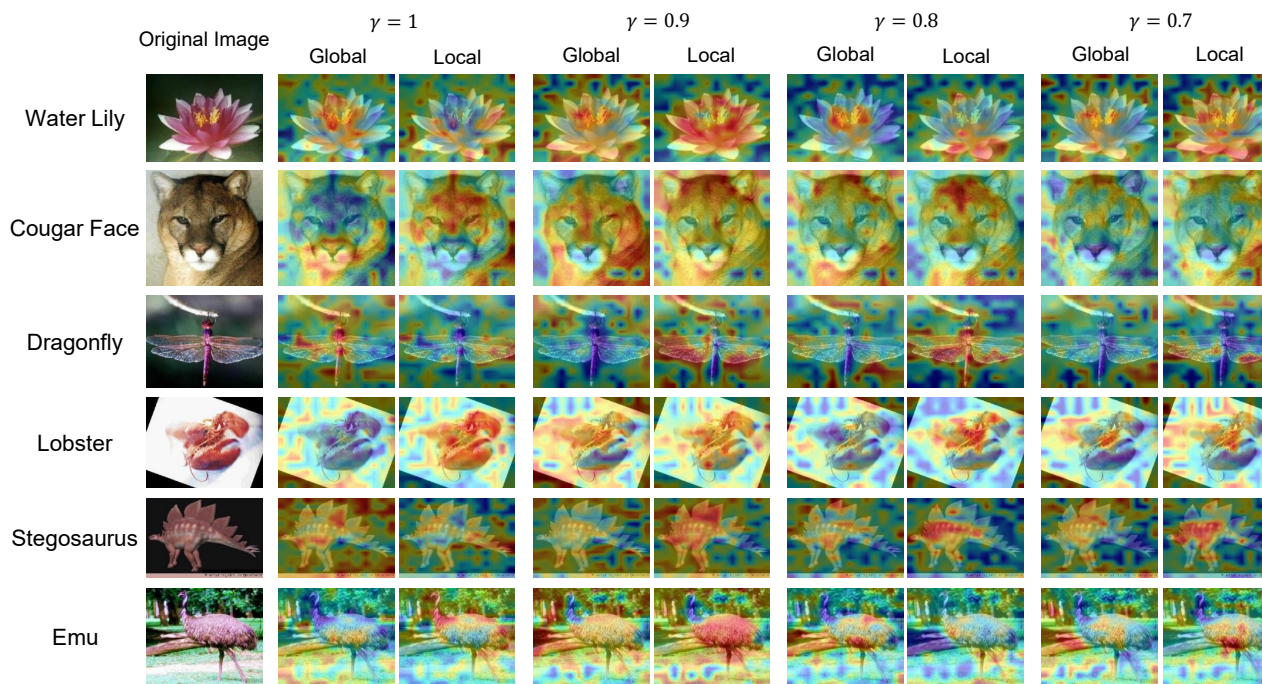
usiv

D.2. T-SNE Projection of Prompts.

To examine how the learned prompts form a meaningful representation across the client space, we employed the t-SNE algorithm [66] to project prompts onto a 2D plane. Following [35, 57], we divided CIFAR-100 dataset into 100 clients. In detail, each coarse label was assigned to five clients, and the corresponding fine labels were uniformly distributed among those



(a) Artifacts in Caltech101 dataset.



(b) Organisms in Caltech101 dataset.

Figure A4. Heatmaps of transport plans related to global and local prompts of FedOTP with different γ in Caltech101 dataset. “Global” denotes the transport plans related to global prompts and “Local” refers to local prompts.

selected clients. After training these clients with FedOTP, we visualized their local prompts obtained from local training, and we used different colors to represent various coarse labels. As shown in Figure A5, local prompts from clients with the same coarse label are clustered together and positioned far from those with different coarse labels. These results further illustrate that in FedOTP, the acquired local prompts are tailored to capture client-specific category characteristics.

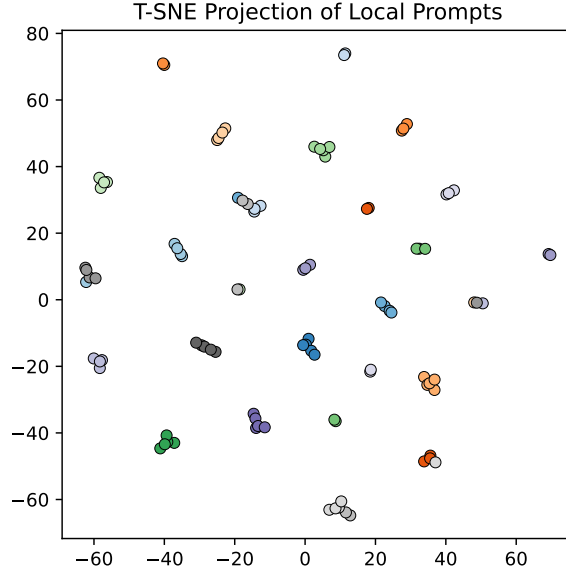


Figure A5. T-SNE projection of local prompts from FedOTP in CIFAR-100 dataset.

E. Generalization Bound

E.1. Key Lemmas

Lemma 1 (McDiarmid’s Inequality [49]) *Let X_1, \dots, X_n be independent random variables, where X_i has range \mathcal{X}_i . Let $g : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \rightarrow \mathbb{R}$ be any function with the (a_1, \dots, a_n) -bounded difference property: for every $i = 1, \dots, n$ and $x_1, \dots, x_n, x'_i \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$, we have*

$$\sup_{x_i \in \mathcal{X}_i} |g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x'_i, \dots, x_n)| \leq a_i. \quad (13)$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}[g(X_1, \dots, X_n) - \mathbb{E}[g(X_1, \dots, X_n)] \geq \varepsilon] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n a_i^2}\right). \quad (14)$$

Lemma 2 (Rademacher Complexity [49]) *Given a space \mathcal{B} and a fixed distribution D_B , let $\{b_1, \dots, b_m\}$ be a set of examples drawn i.i.d. from D_B . Let \mathcal{F} be a class of functions $f : \mathcal{B} \rightarrow \mathbb{R}$, and the Rademacher Complexity of \mathcal{F} is defined as follows:*

$$\mathfrak{R}_{D_B}(\mathcal{B}) = \frac{1}{m} \mathbb{E}_\sigma \left[\sup_{b \in \mathcal{B}} \sum_{i=1}^m \sigma_i b_i \right]. \quad (15)$$

where $\sigma_1, \dots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$.

E.2. Proof of Theorem 1

Proof: For the left side of Theorem 1, we have

$$\begin{aligned}
& \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&= \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) + \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&\leq \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) \right) \right| + \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right|.
\end{aligned} \tag{16}$$

The objective function is partitioned into two components, and we will bound each of them independently. Concerning the first part in Eq. (16), assuming Assumptions 1 and 2 hold, we obtain

$$\begin{aligned}
& \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&\leq \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, P_{l,i}^*) + \mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, P_{l,i}^*) - \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&\leq \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, P_{l,i}^*) \right) \right| + \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, P_{l,i}^*) - \mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&\leq \sum_{i=1}^N \frac{m_i}{M} \mathbb{E}_{(x_i^j, y_i^j) \in \mathcal{D}_i} \left| \ell(f(\hat{P}_g, \hat{P}_{l,i}; x_i^j), y_i^j) - \ell(f(\hat{P}_g, P_{l,i}^*; x_i^j), y_i^j) \right| \\
&+ \sum_{i=1}^N \frac{m_i}{M} \mathbb{E}_{(x_i^j, y_i^j) \in \mathcal{D}_i} \left| \ell(f(\hat{P}_g, P_{l,i}^*; x_i^j), y_i^j) - \ell(f(P_g^*, P_{l,i}^*; x_i^j), y_i^j) \right| \\
&\leq \sum_{i=1}^N \frac{m_i}{M} L \left(\|f(\hat{P}_g, \hat{P}_{l,i}) - f(\hat{P}_g, P_{l,i}^*)\| + \|f(\hat{P}_g, P_{l,i}^*) - f(P_g^*, P_{l,i}^*)\| \right) \\
&\leq \sum_{i=1}^N \frac{m_i}{M} \left(LL_g \|\hat{P}_{l,i} - P_{l,i}^*\| + LL_{l,i} \|\hat{P}_g - P_g^*\| \right) \\
&\leq LL_g A_g + L \sum_{i=1}^N \frac{m_i}{M} L_{l,i} A_{l,i} \leq LL_g A_g + L \left(\sum_{i=1}^N \frac{m_i}{M} \right) \left(\sum_{i=1}^N L_{l,i} A_{l,i} \right) \\
&\leq LL_g A_g + L \sqrt{\left(\sum_{i=1}^N L_{l,i}^2 \right) \left(\sum_{i=1}^N A_{l,i}^2 \right)}.
\end{aligned} \tag{17}$$

For the second part in Eq. (16), replacing $g(\cdot)$ with $\sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right)$ in Lemma 1, and setting $\delta = \exp(-2\varepsilon^2 / \sum_{i=1}^N a_i^2)$, with a probability at least $1 - \delta$, the following inequality holds,

$$\begin{aligned}
& \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| \\
&\leq \mathbb{E} \left[\sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right] + \sqrt{\frac{M}{2} \log \frac{N}{\delta}}.
\end{aligned} \tag{18}$$

Utilizing Lemma 2 and the results in [46], we can get

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right] &\leq \sum_{i=1}^N \frac{m_i}{M} \mathfrak{R}_{\mathcal{D}_i}(\mathcal{H}) \\ &\leq \sum_{i=1}^N \frac{m_i}{M} \sqrt{\frac{dN}{m_i} \log \frac{em_i}{d}} \leq \sum_{i=1}^N \frac{m_i}{M} \sqrt{\frac{dN}{m_i} \log \frac{eM}{d}} \leq \sqrt{\frac{dN}{M} \log \frac{eM}{d}}. \end{aligned} \quad (19)$$

Combining the results in Eq. (18) and Eq. (19), we can get

$$\left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(P_g^*, P_{l,i}^*) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| \leq \sqrt{\frac{M}{2} \log \frac{N}{\delta}} + \sqrt{\frac{dN}{M} \log \frac{eM}{d}}. \quad (20)$$

Summarizing the results above, we can obtain

$$\begin{aligned} \left| \sum_{i=1}^N \frac{m_i}{M} \left(\mathcal{L}_{\hat{\mathcal{D}}_i}(\hat{P}_g, \hat{P}_{l,i}) - \mathcal{L}_{\mathcal{D}_i}(P_g^*, P_{l,i}^*) \right) \right| &\leq \sqrt{\frac{M}{2} \log \frac{N}{\delta}} + \sqrt{\frac{dN}{M} \log \frac{eM}{d}} + LL_g A_g + L \sqrt{\left(\sum_{i=1}^N L_{l,i}^2 \right) \left(\sum_{i=1}^N A_{l,i}^2 \right)} \\ &= \sqrt{\frac{M}{2} \log \frac{N}{\delta}} + \sqrt{\frac{dN}{M} \log \frac{eM}{d}} + L(L_g A_g + L_l A_l), \end{aligned} \quad (21)$$

where we denote $L_l = \sqrt{\sum_{i=1}^N L_{l,i}^2}$ and $A_l = \sqrt{\sum_{i=1}^N A_{l,i}^2}$ for simplicity. ■