Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Fairness-guided federated training for generalization and personalization in cross-silo federated learning*

Ruipeng ZHANG[2,3], Ziqing FAN[2,3], Jiangchao YAO[2,3], Ya ZHANG[†‡1,3], Yanfeng WANG[†‡1,3]

[1]*School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China*

[2]*Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China*

[3]*Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China*

[†]E-mail: ya_zhang@sjtu.edu.cn; wangyanfeng622@sjtu.edu.cn

Received Apr. 12, 2024; Revision accepted May 14, 2024; Crosschecked Nov. 19, 2024; Published online Dec. 16, 2024

Abstract: Cross-silo federated learning (FL), which benefits from relatively abundant data and rich computing power, is drawing increasing focus due to the significant transformations that foundation models (FMs) are instigating in the artificial intelligence field. The intensified data heterogeneity issue of this area, unlike that in cross-device FL, is caused mainly by substantial data volumes and distribution shifts across clients, which requires algorithms to comprehensively consider the personalization and generalization balance. In this paper, we aim to address the objective of generalized and personalized federated learning (GPFL) by enhancing the global model's cross-domain generalization capabilities and simultaneously improving the personalization performance of local training clients. By investigating the fairness of performance distribution within the federation system, we explore a new connection between generalization gap and aggregation weights established in previous studies, culminating in the fairness-guided federated training for generalization and personalization (FFT-GP) approach. FFT-GP integrates a fairness-aware aggregation (FAA) approach to minimize the generalization gap variance among training clients and a meta-learning strategy that aligns local training with the global model's feature distribution, thereby balancing generalization and personalization. Our extensive experimental results demonstrate FFT-GP's superior efficacy compared to existing models, showcasing its potential to enhance FL systems across a variety of practical scenarios.

Key words: Generalized and personalized federated learning; Performance distribution fairness; Domain shift
https://doi.org/10.1631/FITEE.2400279 **CLC number:** TP391.4

1 Introduction

Federated learning (FL) has recently emerged as a prominent privacy-preserving paradigm for collaborative learning on distributed data (McMahan et al., 2017) in data-sensitive domains such as healthcare (Haque et al., 2020; Rieke et al., 2020; Xu A et al., 2022) and finance (Kairouz et al., 2021; Zhu et al., 2021). However, existing FL (Zhao et al., 2018; Karimireddy et al., 2020; Li T et al., 2020c; Li X et al., 2020; Wang et al., 2020) predominantly focuses on the cross-device scenario, characterized by numerous clients each possessing limited data, computing power, and communication capabilities. Consequently, data heterogeneity issues are often observed in data across all clients whose distributions follow a uniform meta simplex (Zhao et al., 2018; Li X et al., 2020). Algorithms targeting data heterogeneity have solely concentrated on convergence on the global distribution (Karimireddy et al., 2020; Li

‡ Corresponding authors

* Project supported by the National Key R&D Program of China (No. 2022ZD0160702), the STCSM (Nos. 22511106101, 18DZ2270700, and 21DZ1100-100), the 111 Plan (No. BP0719010), and the State Key Laboratory of UHD Video and Audio Production and Presentation

ORCID: Ruipeng ZHANG, https://orcid.org/0000-0002-4372-4987; Ziqing FAN, https://orcid.org/0009-0009-1459-3250; Jiangchao YAO, https://orcid.org/0000-0001-6115-5194; Ya ZHANG, https://orcid.org/0000-0002-5390-9053; Yanfeng WANG, https://orcid.org/0000-0002-3196-2347

T et al., 2020c) or personalization on the training clients (Smith et al., 2017; Arivazhagan et al., 2019; Oh et al., 2022), not both.

In recent years, the cross-silo scenario (Huang YT et al., 2021; Huang C et al., 2022) has garnered increased attention, with notable applications in medical image analysis (Xu A et al., 2022) and autonomous driving (Chu et al., 2021; Liu KZ et al., 2022), among others (du Terrail et al., 2022). Unlike the cross-device scenario, the cross-silo framework involves significant data volumes at each client, with each one possessing independent data distributions and sufficient computational and communication resources, as illustrated in Fig. 1. Here, the primary form of data heterogeneity transitions to domain or distribution shifts (Khosla et al., 2012; Cohen et al., 2020; Zhang HR et al., 2021) among clients, necessitating a dual focus on enhancing the generalization of the global model and the personalization of local models. This paradigm is thus also called generalized and personalized federated learning (GPFL) (Jiang et al., 2023; Lu et al., 2023; Zhang RP et al., 2023b).

Meanwhile, the substantial computational power and data abundance in the cross-silo setting pave the way for adapting foundation models (FMs) within the FL framework. FMs like GPT-4 (Achiam et al., 2023) and contrastive language-image pre-training (CLIP) (Radford et al., 2021) have marked significant advancements, revolutionizing artificial intelligence (AI) research and applications. However, FMs are trained solely with large-scale open-source data available on the Internet (Schuhmann et al., 2022). Given that a significant portion of task-relevant data cannot be incorporated into the FM pre-training process due to privacy concerns (Lyu et al., 2023), the importance of FL becomes particularly apparent (Zhuang et al., 2023). FedCLIP and PromptFL have integrated the parameter-efficient fine-tuning (PEFT) (Ding et al., 2023) on top of CLIP into FL. Subsequently, Fed-DPT (Wei et al., 2023) and FedAPT (Su et al., 2024) enhance personalized performance of PromptFL under domain shifts among training clients for domain adaptation. Nevertheless, no CLIP-based FL algorithms have been specifically designed to address the domain shift challenge in the cross-silo scenario for the GPFL objective.

In this paper, we aim to achieve the GPFL objective under both traditional and CLIP-based federated training by promoting fairness and employing a meta-learning framework to personalize local training. In generalization adjustment (GA) (Zhang RP et al., 2023a), altering the aggregation weights for a single client can impact the global model's data distribution performance. Inspired by GA, we discover correlation between the local model's weights and the

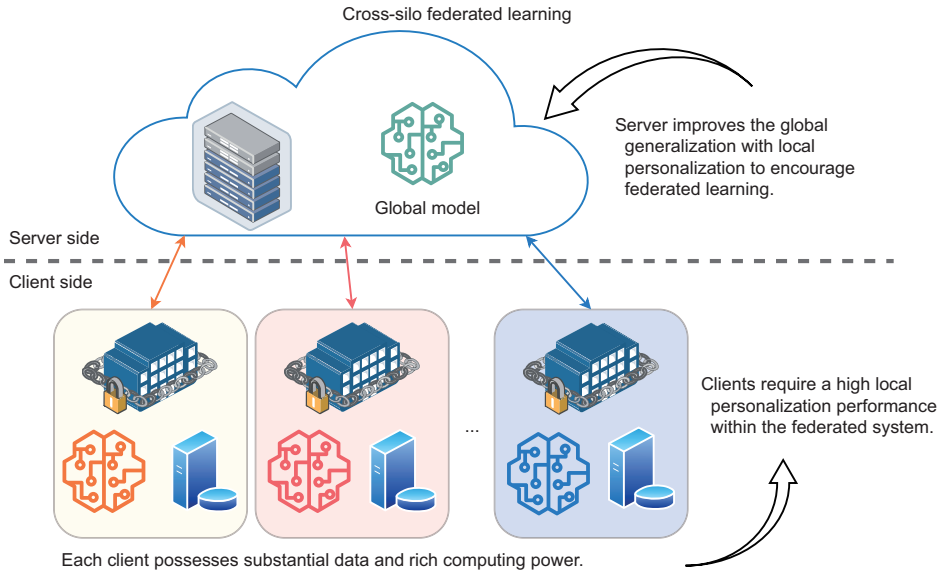

Fig. 1 Illustration of cross-silo FL. It has fewer clients than the cross-device FL, but each client has substantial data and computational resources. From the client perspective, FL needs to enhance each client's personalization performance while maintaining fairness among clients. On the server side, advancing FL requires improving the generalization capability of the global model to foster collaboration. FL: federated learning

performance of the aggregated global model for its client, uncovering an explicit linear relationship between changes in weights and the corresponding loss value. Leveraging this relationship, we introduce an optimal aggregation weight calculation formula to minimize the variance of generalization gaps among training clients, thereby improving performance distribution fairness (Shi et al., 2024). Based on the above observation, we propose fairness-guided federated training for generalization and personalization (FFT-GP). We initially devise a fairness-aware aggregation (FAA) method for updating the global model on the server side, inspired by the linear relationship. Furthermore, we align the feature distribution between the global and local models during local training, employing a meta-learning strategy to balance personalization and generalization. Our main contributions are as follows:

1. We first address the FL problem that balances generalization and personalization in the cross-silo scenario from the fairness perspective. By identifying an explicit linear relationship between aggregation weights and the corresponding client performance, we propose a new method of deriving aggregation weights that maximizes performance distribution fairness at each round.

2. We introduce the FFT-GP, featuring FAA on the server side, where optimal weights for each round are determined by estimating linear correlation coefficients on each client. We constrain the alignment between the global model and local models through meta-learning to achieve balanced generalization and personalization (BGP) in local training.

3. Extensive experiments are conducted under both traditional and CLIP-based federated training. The results show that our FFT-GP algorithm exceeds current state-of-the-art methods on both global model generalization and local model personalization performances. Moreover, through fairness analysis and visualization, we further substantiate the algorithm's effectiveness.

2 Related works

2.1 FL with non-IID data

FL has emerged as a compelling paradigm for multi-site data collaboration, offering significant advantages in communication efficiency and privacy preservation (McMahan et al., 2017; Li T et al., 2020b; Kairouz et al., 2021; Fan et al., 2022). FedAvg (McMahan et al., 2017), the pioneering FL algorithm, adeptly balances privacy preservation and collaborative training by transmitting models rather than data. However, FedAvg presupposes that data are independent and identically distributed (IID). In real scenarios, data across clients are frequently non-independent and non-identically distributed (non-IID), a phenomenon referred to as data heterogeneity (Zhao et al., 2018; Li X et al., 2020; Huang YT et al., 2021; Zhu et al., 2021).

Existing studies tackle the data heterogeneity challenge via two principal strategies. The first strategy enhances the generalizability of the global model, termed generalized federated learning (GFL), aiming at rapid and consistent convergence. FedProx (Li T et al., 2020c) and SCAFFOLD (Karimireddy et al., 2020) enhance global model convergence by incorporating consistency constraints during local training. In contrast, FedNova (Wang et al., 2020) and FedCSA (Ma et al., 2021) improve the global model's generalizability by adjusting weights to reflect variations in dataset sizes or client category distributions during aggregation. The second strategy deviates from the global model, and focuses on personalized local model training aligned with each client's data distribution, referred to as personalized federated learning (PFL). Algorithms like FedMTL (Smith et al., 2017) and Ditto (Li T et al., 2021) view personalization for each client as a distinct task in multi-task learning, with federation aggregation imposing commonality constraints across tasks. Conversely, FedPer (Arivazhagan et al., 2019), FedRep (Collins et al., 2021), and FedBABU (Oh et al., 2022) share certain network layer parameters while others remain personalized.

2.2 Cross-silo FL

Early research focused primarily on the cross-device scenarios (McMahan et al., 2017; Wang et al., 2020; Kairouz et al., 2021; Zhu et al., 2021), where clients possess limited data and computational resources and face connectivity challenges (Zhao et al., 2018; Li X et al., 2020). Despite the vast number of clients, algorithmic designs often addressed only a single facet of generalized or personalized FL.

However, in this paper, we focus on the cross-silo scenario (Huang YT et al., 2021; Huang C et al.,

2022), characterized by fewer clients but with sufficient data and computational capabilities per client. Cross-silo FL is particularly advantageous for applications in sectors like healthcare (Rieke et al., 2020; Xu A et al., 2022) and finance (Huang C et al., 2022), where it aligns well with the inherent requirements of these domains. In this case, data heterogeneity presents a more complex challenge, particularly in domain or distribution shifts (Yuan HL et al., 2022).

Federated domain generalization (FedDG) (Liu QD et al., 2021) was proposed to handle the domain shift problem and aims to improve the generalization of the global model to new unseen clients in the presence of domain shifts across clients. ELCFS (Liu QD et al., 2021) first defines the FedDG and proposes a cross-domain data augmentation method during local training for FL on medical imaging. FADH (Xu QW et al., 2024) introduces a novel algorithm that leverages adversarial sample generation, drawing inspiration from the ELCFS concept. Concurrently, FedSR (Nguyen et al., 2022) incorporates domain-invariant feature constraints into local training, and GA (Zhang RP et al., 2023a) reevaluates the global aggregation for FedDG. CSAC (Yuan JK et al., 2023) unifies multi-source semantic learning and alignment collaboratively by repeatedly alternating semantic aggregation and calibration.

In the cross-silo FL, GRACE (Zhang RP et al., 2023b) and IOP-FL (Jiang et al., 2023) introduce the GPFL framework. This initiative is driven by two key factors: first, the data and computational resources of each client facilitate the simultaneous training of global and personalized local models; second, to effectively establish a federation, it is essential to concurrently optimize the global model's generalization and achieve high performance on individual training clients. Notably, GRACE (Zhang RP et al., 2023b) proposes to correct the updated gradients on both global and local models during federated training, while IOP-FL (Jiang et al., 2023) proposes adaptive model aggregation to accommodate new client integration and facilitate test-time adjustment. In this paper, we endeavor to achieve the objectives of GPFL more effectively, emphasizing the importance of fairness on performance distribution.

2.3 FL with FMs

AI is undergoing a significant paradigm shift with the emergence of FMs like GPT-4 (Achiam et al., 2023), CLIP (Radford et al., 2021), and segment anything model (SAM) (Kirillov et al., 2023). These models are trained on extensive datasets using self-supervision at scale, enabling adaptation to a multitude of downstream tasks. Recent advancements in large-scale vision-language pre-training have spurred the development of CLIP-based FL methods. Given that the FL application scenarios typically impose stringent privacy protection requirements, using FL during the pre-training of FMs proves challenging. Consequently, there is a greater emphasis on the downstream adaptation leveraging pre-trained FM models.

The core idea is to create a federated version of PEFT (Ding et al., 2023), which can fully use the powerful capabilities of FMs (Zhuang et al., 2023). FedCLIP (Lu et al., 2023) pioneers the use of a pre-trained CLIP model for each client, implementing federated training on an additional adapter layer following the image encoder. Similarly, PromptFL (Guo et al., 2024) introduces an innovative approach by incorporating a prompt learning technique (Zhou et al., 2022) into the CLIP text encoder. Expanding on this, FedAPT (Su et al., 2024) and Fed-DPT (Wei et al., 2023) further enhance the prompt learning strategy, focusing on domain adaptation to refine its effectiveness.

2.4 Fairness in FL

During federated model training, variations in data quality, quantity, and local client resources lead to differing contributions to the final FL model. This diversity introduces fairness as a critical research challenge within FL (Zeng et al., 2021; Shi et al., 2024). Fairness concerns arise at various stages of FL, including model optimization, contribution evaluation, client selection, and incentive mechanisms. In this study, we specifically examine how enhancing model optimization fairness affects the global model's generalizability in the cross-silo scenario. AFL (Mohri et al., 2019), an early work in this area, sought to establish fairness by preventing model overfitting to any single client to the detriment of others. Following this, q-FFL (Li T et al., 2020a) was proposed, drawing inspiration from fair resource allocation strategies to ensure more uniform accuracy across FL devices. Additionally, AdaFFL (Cong et al., 2023) introduces a dynamic adjustment of q-FFL's fairness coefficient based on local

model updates. Similarly, Cui et al. (2021) proposed a gradient-based procedure to enforce both algorithmic fairness and performance consistency, aiming to achieve a Pareto-optimal solution. Moreover, Zhang FD et al. (2024) introduced a unified framework, FedUFO, which uses distributional robust optimization and a unified uncertainty set to enhance performance consistency and training efficacy for digital healthcare tasks.

Although these methods evaluate fairness in terms of accuracy variance, the presence of domain shifts suggests that the generalization gap (a concept introduced in GA (Zhang RP et al., 2023a)) serves as a more appropriate fairness metric. Our proposed approach focuses on improving performance distribution fairness to enhance generalization within the GPFL framework.

3 FFT-GP

In this section, we will first delineate the optimization objective of GPFL. Subsequently, we will elaborate on the motivation behind and the implementation details of our proposed algorithm.

3.1 Preliminaries

Denote the set of datasets of all clients as $\mathcal{D}=\{D_1,D_2,\cdots\}$ and the sampled counterpart for training as $\mathcal{S}=\{D_1,D_2,\cdots,D_M\}$, where M is the number of training clients (or domains, with each client representing one domain). A sample pair from D_i is denoted as $(\boldsymbol{x}^i,y^i)$, and the loss function $\mathcal{L}$ measures the distance between the model's prediction $f(\boldsymbol{x};w)$, parameterized by w, and label y. The notations used throughout this paper are summarized in Table 1 for clarity. Furthermore, given a sampled counterpart $D_i=\{\boldsymbol{x}_j^i,y_j^i\}_{j=1}^{N_i}, N_i=|D_i|$, the empirical risk on client (or domain) i is defined as

$$\mathbb{E}_{D_i}(w)=\frac{1}{N_i}\sum_{j=1}^{N_i}\mathcal{L}\left(f(\boldsymbol{x}_j^i;w),y_j^i\right). \tag{1}$$

GPFL has two ideal objectives: from the perspective of global generalization, the resulting global model is expected to exhibit excellent performance on $\mathcal{D}$ despite domain shifts; from the viewpoint of training clients, each client aims to minimize the expected loss on their local data distribution. In practice, we typically have sampled domains and their corresponding data points $\{\boldsymbol{x}_j^i,y_j^i\}_{j=1}^{N_i}$ in each domain i. Thus, FedAvg optimizes the following objective:

$$\begin{aligned}\min_w\ \mathbb{E}_{\mathcal{S}}(w)&=\sum_{i=1}^{M}p_i\mathbb{E}_{D_i}(w)\\&=\sum_{i=1}^{M}\frac{p_i}{N_i}\left(\sum_{j=1}^{N_i}\mathcal{L}\left(f(\boldsymbol{x}_j^i;w),y_j^i\right)\right),\end{aligned} \tag{2}$$

where $p_i=\frac{N_i}{\sum_{m=1}^{M}N_m}$ and we denote $w_{\mathrm{g}}=\arg\min_w\mathbb{E}_{\mathcal{S}}(w)$ as the parameter of global model. In federated training, because the data are stored on each client, obtaining the optimized global model w_{g} directly from the overall training set $\mathcal{S}$ is not feasible. Actually, in FedAvg w_{g} is derived by aggregating the local models' parameters $\{w_i\}_{i=1}^{M}$, which are uploaded after local training at each training client, using weights p_i: $w_{\mathrm{g}}=\sum_{i=1}^{M}p_iw_i$.

Therefore, GPFL aims to minimize both the global model w_{g} with respect to $\mathcal{S}$ and the local models w_i with respect to D_i, where $i=1,2,\cdots,M$:

$$\text{Global}:\min_{w_{\mathrm{g}}}\mathbb{E}_{\mathcal{S}}(w_{\mathrm{g}});\ \text{Local}:\begin{cases}\min_{w_1}\ \mathbb{E}_{D_1}(w_1),\\ \min_{w_2}\ \mathbb{E}_{D_2}(w_2),\\ \quad\cdots\\ \min_{w_M}\ \mathbb{E}_{D_M}(w_M).\end{cases} \tag{3}$$

3.2 Motivation

Numerous studies to address the data heterogeneity issue in FL have highlighted that FedAvg operates by assuming the IID hypothesis for all data from a global training perspective (Zhao et al.,

Table 1 Notations used in this paper

Notation	Description
$D_i=\{\boldsymbol{x}_j^i,y_j^i\}_{j=1}^{N_i}$	Dataset of client i with data $\boldsymbol{x}$ and label y
$\mathcal{S}=\{D_i\}_{i=1}^{M}$	Overall training set with M clients
$p_i=\frac{N_i}{\sum_{m=1}^{M}N_m}$	Weight for client i of FedAvg
a_i	Learnable weight for client i
$f(\boldsymbol{x};w)$	Model f with parameter w and input $\boldsymbol{x}$
w_{g}	Parameter of the global model
w_i, w_i^*	Local model parameters and the optimal values
$\mathcal{L}(f(\boldsymbol{x}^i;w),y^i)$	Performance on client i with loss $\mathcal{L}$
$\mathbb{E}_{D_i}(w)$	Empirical risk on client i
$G_{D_i}(w_{\mathrm{g}})$, G_i	Generalization gap on client i (abbr. G_i)
$t\in\{1,2,\cdots,T\}$	The index of FL round
K_i^t	Linear slope on client i at round t
η	Learning rate
λ	Weight of alignment constraint in BGP

FL: federated learning; BGP: balanced generalization and personalization

2018; Li X et al., 2020; Ma et al., 2021; Zhang RP et al., 2023a, 2023b). Consequently, employing fixed aggregation weights is effective mainly in class-imbalance experiments on datasets such as CIFAR-10/100 and Tiny-ImageNet (Fan et al., 2022, 2023a, 2023b; Zhang FD et al., 2023).

However, the use of fixed weights results in a static relationship between the global objective and the local objectives on each client. In the cross-silo scenario, where domain shifts occur across different clients, FedAvg's fixed weights prevent it from achieving an optimal balance between global generalization and local personalization in GPFL.

Building on the GA algorithm (Zhang RP et al., 2023a), our goal remains to enhance out-of-domain generalization by improving the fairness of the global model's performance distribution, taking into account the generalization gap post-aggregation. Unlike GA, our study identifies an explicit linear relationship between changes in weights and shifts in the generalization gap during the aggregation process.

In GA, the generalization gap $G_{D_i}(w_{\mathrm{g}})$ is employed to assess the bias of the global model w_{g} towards distinct clients, using its variance $\mathrm{Var}\{G_{D_i}(w_{\mathrm{g}})\}_{i=1}^{M}$ as an indicator of performance distribution fairness. A lower variance signifies that the global model demonstrates more uniform generalization performance across various training clients, suggesting enhanced generalization. Formally, the definition of $G_{D_i}(w_{\mathrm{g}})$ is as follows:

$$G_{D_i}(w_{\mathrm{g}}) = G_{D_i}\left(\sum_{j=1}^{M} a_j w_j^*\right) = \mathbb{E}_{D_i}(w_{\mathrm{g}}) - \mathbb{E}_{D_i}(w_i^*), \tag{4}$$

where w_i^* represents the local optimum on domain D_i, a_j denotes the aggregation weight for w_j^*, and $w_{\mathrm{g}} = \sum_{j=1}^{M} a_j w_j^*$. In practice, w_i acquired at the conclusion of each local training round is regarded as an approximation of w_i^*, having converged on the local data distribution.

The former definition of performance distribution was proposed by Li T et al. (2020a), which uses the variance of the performance distribution as a measure of uniformity:

$$\mathrm{Var}\{\mathbb{E}_{D_i}(w_{\mathrm{g}})\}_{i=1}^{M} = \sum_{i=1}^{M}\sum_{j=1}^{M}(\mathbb{E}_{D_i}(w_{\mathrm{g}}) - \mathbb{E}_{D_j}(w_{\mathrm{g}}))^2. \tag{5}$$

We propose that using the absolute value of performance to estimate fairness is not accurate enough when the data distribution is inconsistent across different clients and the task difficulty is not uniform. Forcibly constraining the performance to be the same will suppress the performance of clients with relatively fast convergence, which is not conducive to the formation of the federation. Therefore, in this paper, we use the variance of generalization gaps as the metric for estimating fairness:

$$\begin{aligned} &\mathrm{Var}\{G_{D_i}(w_{\mathrm{g}})\}_{i=1}^{M} \\ =&\sum_{i=1}^{M}\sum_{j=1}^{M}(G_{D_i}(w_{\mathrm{g}}) - G_{D_j}(w_{\mathrm{g}}))^2 \\ =&\sum_{i=1}^{M}\sum_{j=1}^{M}((\mathbb{E}_{D_i}(w_{\mathrm{g}}) - \mathbb{E}_{D_i}(w_i^*)) \\ &-(\mathbb{E}_{D_j}(w_{\mathrm{g}}) - \mathbb{E}_{D_j}(w_j^*)))^2. \end{aligned} \tag{6}$$

Here, Eq. (5) can be considered a special case of Eq. (6) when the local optimal performance is the same among clients.

Building on the aforementioned $G_{D_i}(w_{\mathrm{g}})$ definition, the fairness-aware global objective $\mathbb{E}_{\mathcal{S}}^{\mathrm{fairness}}(w_{\mathrm{g}})$ is formulated as

$$\begin{aligned} &\min_{w_1,w_2,\cdots,w_M,\boldsymbol{a}} \mathbb{E}_{\mathcal{S}}^{\mathrm{fairness}}(w_{\mathrm{g}}) \\ =&\sum_{i=1}^{M} a_i \mathbb{E}_{D_i}(w_i) + \beta \mathrm{Var}\{G_{D_i}(w_{\mathrm{g}})\}_{i=1}^{M} \\ \text{s.t. }&\sum_{i=1}^{M} a_i = 1, w_{\mathrm{g}} = \sum_{i=1}^{M} a_i w_i, \forall i, a_i \geq 0. \end{aligned} \tag{7}$$

Here, we denote the learnable client/domain weights by $\boldsymbol{a} = (a_1, a_2, \cdots, a_M)$, and $\beta \in [0, \infty)$ modulates the equilibrium between minimizing global risk and promoting fairness among generalization gaps, with $\beta = 0$ reverting to the FedAvg algorithm and $\beta \to \infty$ equating the generalization gaps exclusively. To simplify the notation, we abbreviate $G_{D_i}(w_{\mathrm{g}})$ as G_i, and $G_{D_i}(w_{\mathrm{g}}^t)$ in the t^{th} round is represented as G_i^t.

GA proposes adjusting the magnitude of a_i to influence G_i, aiming to reduce the variance in $\{G_i\}_{i=1}^{M}$ through a weight adjustment strategy based on a predetermined step size. However, this approach necessitates the prior specification of a step size hyper-parameter, whose excessive or insufficient magnitude can impact the aggregation outcome and fails to guarantee the minimization of G_i's variance through adjusted weights.

We further explore the relationship between a_i and G_i. As depicted in Fig. 2, we examine the linearity between Δa and ΔG for each training client across various benchmarks. We specifically adjust the aggregation weights to influence the proportion of the local model on an individual client and then calculate the generalization gap of the aggregated global model for that client. Our empirical analysis across multiple benchmarks shows a distinct linear correlation between Δa and ΔG for the training clients.

Additionally, we note that the slope of the linear relationship between Δa and ΔG differs among clients due to the domain shifts, which results in considerable variability among the trained local models. These findings, as presented in Fig. 2, highlight the importance of identifying optimal aggregation weights during the aggregation phase. Based on these observations, we introduce an optimal aggregation weight formula aimed at maximizing fairness within the linear relationship.

3.3 FFT-GP algorithm

Here, we introduce in detail our method: FFT-GP, which encompasses FAA for global generalization and BGP for personalizing local training (as shown in the left side of Fig. 3). Then, we will discuss the application of foundation models in federated learning (as shown in the right part of Fig. 3).

3.3.1 FAA

On the cloud server side, the optimization of the global objective in FL is achieved by aggregating the local models uploaded by each client. To realize the global optimization objective (7) of enhanced fairness, we must calculate appropriate aggregation weights to minimize the variance of the

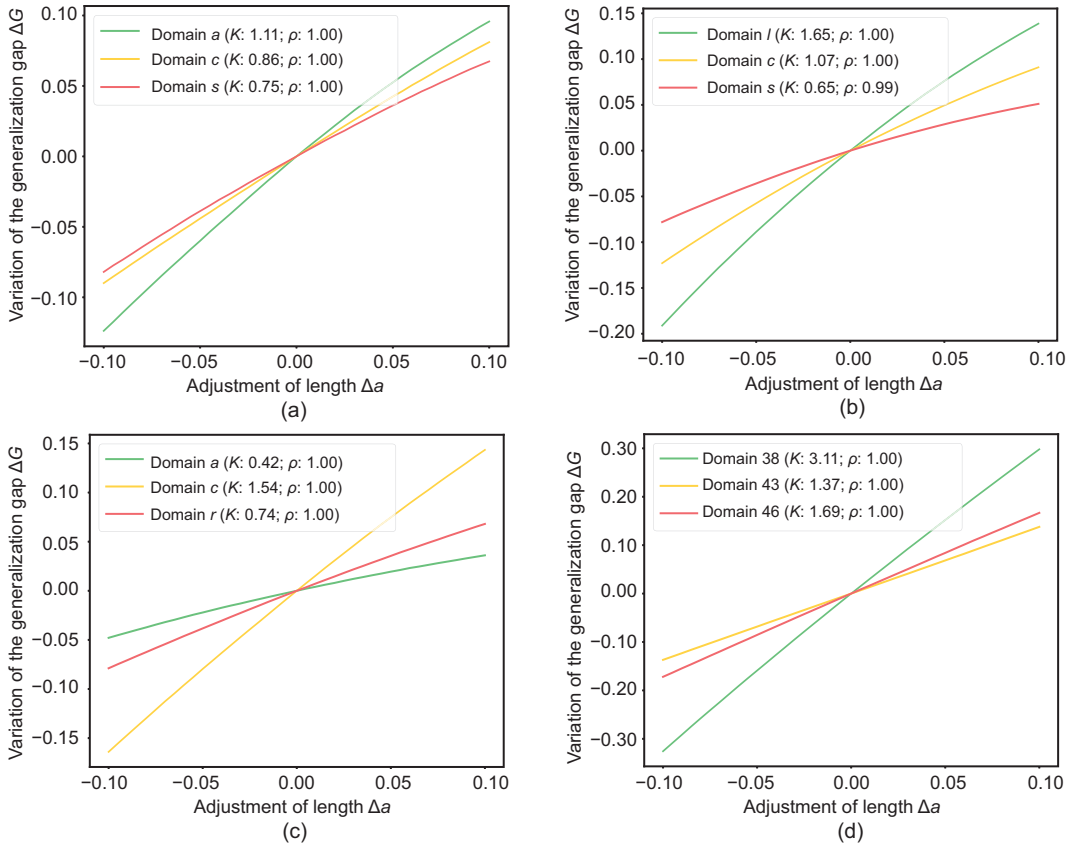

Fig. 2 Relationship between the change in aggregation weights Δa and the change in generalization gaps ΔG across various benchmarks: (a) PACS; (b) VLCS; (c) OfficeHome; (d) TerraInc. The legend in each subplot denotes the linear correlation slope K and the Pearson linear correlation coefficient ρ for Δa and ΔG on each training client (Details of the experimental settings are presented in Section 4.1). References to color refer to the online version of this figure

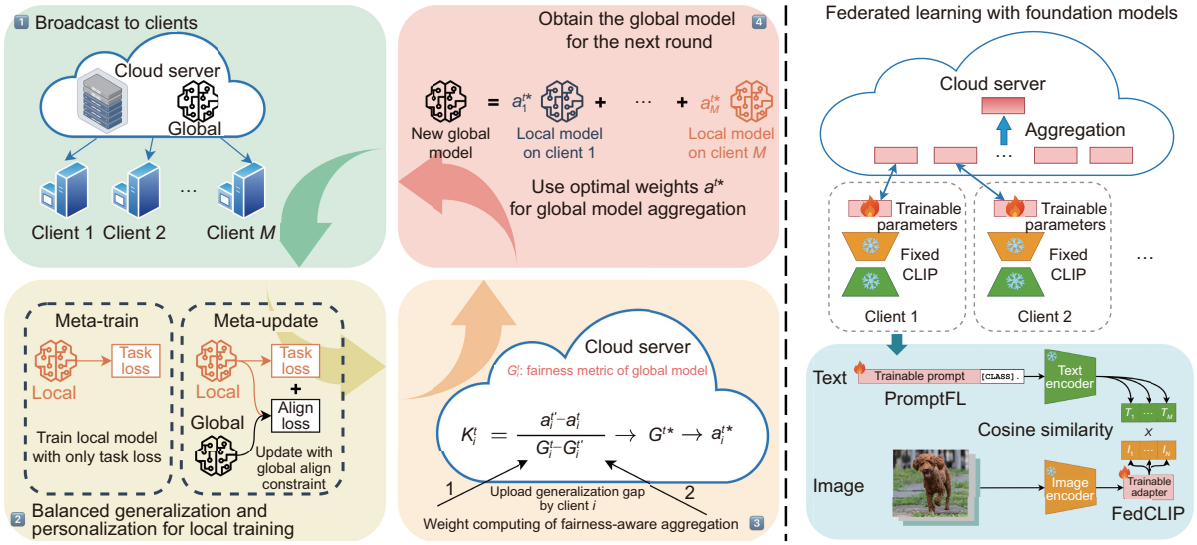

Fig. 3 Left: the training framework of our method, which unfolds in a cyclical sequence of four pivotal steps: (1) broadcasting the global model to each client; (2) executing meta-learning on each training client to enhance both generalization and personalization; (3) adjusting weights in a fairness-aware manner based on the generalization gap reported by each client; (4) aggregating the local models using the revised weights to formulate a new global model. This iterative process continues until the predetermined number of training cycles is reached. Right: illustrations of FL with FMs. The pink rectangles denote the portions of trainable parameters that undergo fine-tuning within FMs. Additionally, the bottom figure demonstrates how these trainable parameters are integrated with the pre-trained CLIP model in both FedCLIP and PromptFL. CLIP: contrastive language-image pre-training; FL: federated learning; FMs: foundation models. References to color refer to the online version of this figure

generalization gap.

In Section 3.2, we identified a new insight into the relationship between the changes in aggregation weights a_i and the changes in generalization gaps G_i, specifically, a linear relationship between Δa_i and ΔG_i. It is crucial to acknowledge that the slope of the change curve between Δa and ΔG varies across different domains, highlighting the domain shift problem among training clients. Under the hypothesis of linear correlation, for the local model w_i^t that has completed its local training in the t^{th} round, if the slope K_i^t of a_i^t and G_i^t for each domain is known, we can calculate the weight vector $\boldsymbol{a}^t = (a_1^t, a_2^t, \cdots, a_M^t)$ that minimizes the variance of $\{G_i\}_{i=1}^M$. This principle is the cornerstone of our FAA. The specific calculation process is detailed below:

At the t^{th} round, for the set of all local models $\{w_1^t, w_2^t, \cdots, w_M^t\}$, we denote the optimal aggregation weight vector as $\boldsymbol{a}^{t*} = (a_1^{t*}, a_2^{t*}, \cdots, a_M^{t*})$ and the adjusted fair global model as $w_{\text{g}}^{t*} = \sum_{i=1}^M a_i^{t*} w_i^t$, where for w_{g}^{t*}, $G_{D_1}(w_{\text{g}}^{t*}) = G_{D_2}(w_{\text{g}}^{t*}) = \cdots = G_{D_M}(w_{\text{g}}^{t*}) = G^{t*}$. For the remaining weight vectors $\boldsymbol{a}^t \neq \boldsymbol{a}^{t*}$, the corresponding globally aggregated model is w_{g}^t, and the generalization gap of w_{g}^t on each client i is G_i^t. Leveraging the linear correlation and the correlation slope K_i for each domain, we can derive the following nonlinear equations:

$$\begin{cases} a_1^{t*} - a_1^t = K_1^t (G_1^t - G^{t*}), \\ a_2^{t*} - a_2^t = K_2^t (G_2^t - G^{t*}), \\ \cdots \\ a_M^{t*} - a_M^t = K_M^t (G_M^t - G^{t*}), \\ \sum_{i=1}^M a_i^t = 1, \\ \sum_{i=1}^M a_i^{t*} = 1. \end{cases} \tag{8}$$

Here we also incorporate the constraint that the sum of all aggregation weights a_i equals 1. It is crucial to observe that for any client i, as the weight a_i^t increases, the corresponding G_i^t decreases, and hence the slope K_i^t in Eq. (8) is consistently ≥ 0, namely $\forall i \in \{1, 2, \cdots, M\}$ and $t \in \{1, 2, \cdots, T\}$, $K_i^t \geq 0$. In Eq. (8), we can use a random aggregation weight vector $\boldsymbol{a}^t$ to obtain the corresponding $\{G_i^t\}_{i=1}^M$. We now know $\{a_i^t\}_{i=1}^M$, $\{G_i^t\}_{i=1}^M$, and $\{K_i^t\}_{i=1}^M$. Then, the value of the minimized generalization gap can be

calculated as

$$G^{t*}=\frac{\sum_{i=1}^{M}K_i^tG_i^t}{\sum_{i=1}^{M}K_i^t}. \tag{9}$$

The corresponding aggregation weights are

$$a_i^{t*}=K_i^t(G_i^t-G^{t*})+a_i^t. \tag{10}$$

Practically, because the relationship between the aggregation weight a_i and the generalization gap G_i is linearly constrained by K_i on client i, K_i can be determined by solving Eq. (11) for $\boldsymbol{a}$ after perturbing $a_i^{t'}$:

$$K_i^t=\frac{a_i^{t'}-a_i^t}{G_i^t-G_i^{t'}}. \tag{11}$$

In our FAA, we initially compute G_i through aggregation using the weight vector from the previous round $\boldsymbol{a}^t=\boldsymbol{a}^{t-1}$, and then apply the GA (Zhang RP et al., 2023a) algorithm for a weight perturbation to acquire $a_i^{t'}$ and the corresponding $G_i^{t'}$. Then, we calculate the slope K_i for each client i. Subsequently, we input these results into Eq. (9) to obtain G^{t*}, and then incorporate G^{t*} into Eq. (10) to determine the optimal aggregation weight $\boldsymbol{a}^{t*}$ for the t^{th} round. The overall aggregation process is shown in Algorithm 1.

Algorithm 1 FAA

Input: Set of all local models $\{w_1^t,w_2^t,\cdots,w_M^t\}$ at round t and the previous weight vector $\boldsymbol{a}^{t-1}$
Output: Global model w_{g}^{t+1} for the next round $t+1$
Initialize $\boldsymbol{a}^t=\boldsymbol{a}^{t-1}$
for each client $i=1$ to M **do**
  Compute G_i through aggregation using $\boldsymbol{a}^t$
  Apply the GA algorithm to perturb weights to obtain $a_i^{t'}$ and the corresponding $G_i^{t'}$
  Calculate the slope $K_i^t=\frac{a_i^{t'}-a_i^t}{G_i^t-G_i^{t'}}$
end for
Compute $G^{t*}=\frac{\sum_{i=1}^{M}K_i^tG_i^t}{\sum_{i=1}^{M}K_i^t}$
for each client $i=1$ to M **do**
  Update weight $a_i^{t*}=K_i^t(G_i^t-G^{t*})+a_i^t$
end for
Global model: $w_{\text{g}}^{t+1}=\sum_{i=1}^{M}a_i^{t*}w_i^t$
return w_{g}^{t+1}

3.3.2 BGP for local training

As illustrated in Fig. 3, a round of federated training commences with the cloud server broadcasting the initial training parameters. In addressing the requisites for GPFL within the cross-silo scenario, we design a meta-learning framework that harmonizes generalization and personalization in the training of local models. PerFedAvg (Fallah et al., 2020) initially introduces meta-learning to local training, aiming to enhance local personalization beyond what is achievable with FedAvg, without focusing on the generalization of the global model. Our meta-learning framework integrates alignment constraints with the global model during local personalization, ensuring that the personalized local models do not compromise the generalization of the aggregated global model. Its fundamental concept is that a more generalizable global model aids local models in better adapting to their respective local distributions, which in turn can offer constructive feedback to the global model through enhanced gradients. The meta-learning framework for local training executes two procedures during each gradient update.

1. Meta-train

Meta-train comprises one step of gradient descent, targeting the task-related loss (cross-entropy loss for classification tasks), thereby personalizing to the local data distribution for each client. We represent $w_i^{t,k}$ as the personalized model parameter at the local update step k for client i in round t, and η as the local learning rate. The initial personalization step is guided only by the task loss:

$$\begin{aligned}(w_i^{t,k})'&=w_i^{t,k}-\eta\nabla\mathcal{L}_i^{\text{task}}(w_i^{t,k})\\&=w_i^{t,k}-\eta\nabla\mathcal{L}(f(\boldsymbol{x}_{\text{tr}}^i;w_i^{t,k}),y_{\text{tr}}^i),\end{aligned} \tag{12}$$

where $(\boldsymbol{x}_{\text{tr}}^i,y_{\text{tr}}^i)\in D_i$ denotes the sampled data and $(w_i^{t,k})'$ is the updated parameter that will be used in the subsequent step.

2. Meta-update

After optimizing the local task objective, a meta-update is performed to virtually assess the updated parameters $(w_i^{t,k})'$ using the held-out meta-test data $(\boldsymbol{x}_{\text{te}}^i,y_{\text{te}}^i)\in D_i$ with a meta-objective $\mathcal{L}^{\text{meta}}$. This involves recalculating the task loss based on the parameters refined in the first step, while integrating a feature alignment constraint with the global model. In the meta-update's loss function, we incorporate a feature alignment regularizer:

$$\begin{aligned}&\mathcal{L}_i^{\text{meta}}(\boldsymbol{x}_{\text{te}}^i,y_{\text{te}}^i;(w_i^{t,k})')\\=&\mathcal{L}(f(\boldsymbol{x}_{\text{te}}^i;(w_i^{t,k})'),y_{\text{te}}^i)\\&+\lambda\mathcal{L}^{\text{align}}(h(\boldsymbol{x}_{\text{te}}^i;\phi_{\text{g}}^t),h(\boldsymbol{x}_{\text{te}}^i;(\phi_i^{t,k})')),\end{aligned} \tag{13}$$

where $h(\cdot;\phi)$ is the feature extractor part of model $f(\cdot;w)$ with ϕ as the corresponding parameter, and

λ is the weight for the alignment loss, typically set to 1.0. The CORAL alignment loss (Sun and Saenko, 2016), which has shown optimal results in DomainBed (Gulrajani and Lopez-Paz, 2021) for domain generalization, is employed here. This process guarantees that the local models, despite undergoing personalization, maintain alignment with the global model in terms of feature distribution. The overall local training process is illustrated in Algorithm 2.

3.4 Using FMs for FL

CLIP is a weakly supervised learning paradigm that integrates visual and language encoders to tackle image recognition challenges. As depicted on the right side of Fig. 3, the CLIP model maps the input image and text through respective encoders to a shared space, subsequently calculating the cosine similarity between the derived image and text features. Image classification is accomplished based on the relative magnitudes of similarity with various texts.

Although the CLIP pre-trained model presents impressive zero-shot generalization, its parameter scale and computational demands have increased significantly compared to prior supervised learning models, thus posing challenges for full parameter fine-tuning within an FL framework. FedCLIP first introduces a GPFL-based approach that necessitates only PEFT of the CLIP model during the local training phase, which reduces the parameter scale being updated and, consequently, diminishes the overall communication and computational burdens of federated training. Specifically, FedCLIP employs the CLIP-Adapter method, depicted in the lower right corner of Fig. 3, integrating a residual module with two linear layers after the CLIP image encoder to tailor the image features. Likewise, PromptFL leverages prompt learning from PEFT, appending trainable prompt vectors before the text encoder to enhance the alignment of image and text features via text feature adaptation. These two types of CLIP-based FL frameworks are illustrated in Fig. 3. Notably, in the PromptFL framework, feature alignment of FFT-GP targets text features instead of image features, because the image encoder component is fixed.

In this study, we implement the traditional federated training framework along with several CLIP-based federated frameworks to verify the effectiveness of the FFT-GP algorithm.

Algorithm 2 BGP for local training

Input: Initial global model w_{g}^{t} at round t
Output: Updated local models $\{w_i^t\}_{i=1}^{M}$

for client $i = 1, 2, \cdots, M$ **do**
  Initialize $w_i^{t,0} \leftarrow w_{\mathrm{g}}^{t}$
  for local step $k = 0, 1, \cdots, K-1$ **do**
    /* meta-train */
    Sample mini-batch $(\boldsymbol{x}_{\mathrm{tr}}^{i}, y_{\mathrm{tr}}^{i}) \in D_i$
    Calculate task-related loss $\mathcal{L}^{\mathrm{task}}$
    Update $(w_i^{t,k})' \leftarrow w_i^{t,k} - \eta \nabla \mathcal{L}_i^{\mathrm{task}}(w_i^{t,k})$
    /* meta-update */
    Sample $(\boldsymbol{x}_{\mathrm{te}}^{i}, y_{\mathrm{te}}^{i})$ for meta-test
    Calculate meta-objective loss
      $\mathcal{L}_i^{\mathrm{meta}} = \mathcal{L}^{\mathrm{task}} + \lambda \mathcal{L}^{\mathrm{align}}$
    Update $w_i^{t,k+1} \leftarrow (w_i^{t,k})' - \eta \nabla \mathcal{L}_i^{\mathrm{meta}}((w_i^{t,k})')$
  end for
  Obtain local model $w_i^t = w_i^{t,K}$
end for

3.5 Cost by FFT-GP

Here we present the additional overhead introduced by using our FFT-GP algorithm. Compared to the baseline method FedAvg, FFT-GP's meta-learning framework in the local training phase doubles the local computational overhead. Additionally, employing the FAA global aggregation method, which necessitates estimating the slope K_i of the linear relationship, results in two extra global model broadcasts, increasing communication overhead and computational costs for client-side evaluations.

It is crucial to recognize that the introduction of the CLIP-based federated training framework significantly reduces the overall transmission and local training update parameters. In the cross-silo scenario, the emphasis of the federated system shifts towards balancing the generalization of the global model and the personalization of local models, rather than merely focusing on computational and communication costs.

If FFT-GP is reduced to the GA's progressive adjustment method, a substantial reduction in overhead can be achieved. In practice, the choice between GA and FAA methods can be adjusted based on real-world conditions, allowing for trade-off between performance and costs.

4 Experiments

In this section, we demonstrate the in-domain personalization and out-of-domain generalization

performances of the FFT-GP algorithm compared with state-of-the-art methods on multiple benchmark datasets. In addition, we compare the fairness of the global model achieved by different aggregation strategies and visualize the feature distribution to show the differences caused by domain shift during training.

4.1 Experiment setups

Here we introduce the datasets and baselines used in the experiments, followed by the implementation details.

4.1.1 Datasets

We first evaluate our proposed method on five widely used FedDG benchmarks for the image classification task, namely, PACS (Li D et al., 2017) (9991 images across four domains), VLCS (Fang et al., 2013), OfficeHome (Venkateswara et al., 2017) (15 588 images across four domains), TerraInc (Beery et al., 2018) (24 788 images across four domains), and DomainNet (Peng et al., 2019) (569 010 images across six domains). We implement a leave-one-domain-out splitting and evaluation strategy (Li D et al., 2017; Gulrajani and Lopez-Paz, 2021; Liu QD et al., 2021; Zhang RP et al., 2023a) for all benchmarks. Specifically, we iteratively select one domain as the unseen client, while using the remaining domains as the source clients for training, with each domain representing one client. The splitting of training and validation sets within each source domain follows the DomainBed benchmark (Gulrajani and Lopez-Paz, 2021; Xu QW et al., 2021; Lu et al., 2023; Zhang RP et al., 2023a) for PACS, VLCS, OfficeHome, TerraInc, and DomainNet, with the entire target domain used for testing. We follow the same divided subsets with 10 popular classes of DomainNet on FedBN (Li XX et al., 2021) and GA (Zhang RP et al., 2023a). As for evaluation, personalization performance is calculated as the mean classification accuracy of local models on each training client's validation set, while generalization performance is defined as the classification accuracy of the global model on the overall left-out unseen domain. The reported results for each benchmark represent the averaged accuracy across all split configurations.

Additionally, we conduct extensive validation on federated benchmarks involving multiple modalities, including the Fed-Prostate segmentation task (Liu QD et al., 2021) for medical magnetic resonance imaging (MRI) images and the Shakespeare (Caldas et al., 2018) prediction task for text data. The Fed-Prostate benchmark is a prostate MRI segmentation task (Liu QD et al., 2021; Jiang et al., 2023) comprising T2-weighted MRI images across six different domains (Litjens et al., 2014; Bloch et al., 2015; Lemaître et al., 2015; Liu QD et al., 2020). The Shakespeare benchmark is derived from *The Complete Works of William Shakespeare* (McMahan et al., 2017), with each speaking role associated with a device. We follow all training and evaluation settings described in Liu QD et al. (2021) for the Fed-Prostate and Li T et al. (2020a) for the Shakespeare.

4.1.2 Baselines

We select several advanced algorithms to deploy in both traditional supervised learning-based federated training, such as FedAvg (McMahan et al., 2017), and CLIP-based federated training frameworks, such as FedCLIP (Lu et al., 2023) and PromptFL (Guo et al., 2024). Here, FedCLIP and PromptFL differ solely in the locations of the trainable parameters, with the parameter training method for both being FedAvg. We select state-of-the-art algorithms from various perspectives and conduct experiments within the three aforementioned frameworks. For the personalization of local models, we choose Ditto (Li T et al., 2021); for the generalization of the global model, we compare FedProx (Li T et al., 2020c) and GA (Zhang RP et al., 2023a); for GPFL, we select PerFedAvg (Fallah et al., 2020) and GRACE (Zhang RP et al., 2023b) for comparison. We also compare the personalization and generalization performances under a fully supervised framework with classical fairness-aware FL methods, including AFL (Mohri et al., 2019) and q-FFL (Li T et al., 2020a).

4.1.3 Implementation details

For the five image classification benchmarks, we follow the protocols used in GA (Zhang RP et al., 2023a) and FedCLIP (Lu et al., 2023) during the local training stage. Specifically, we employ the ImageNet pre-trained ResNet50 (He et al., 2016) for traditional federated training and the

pre-trained CLIP model with ViT-B/16 (Dosovitskiy et al., 2021) image backbone for FedCLIP and PromptFL. To ensure local model convergence within each round's local training phase, we set the number of local epochs E to 5 for ResNet50 and 1 for ViT-B/16, and the number of total communication rounds R to 40 for ResNet50 and 100 for ViT-B/16. For consistency, we maintain a batch size of 16 and a learning rate of 1e-3 during local training in supervised federated training with ResNet50. For FedCLIP and PromptFL, we follow the settings from the original papers (Lu et al., 2023; Guo et al., 2024): the batch size is set to 32, the learning rate for FedCLIP is 5e-5 with a learnable two-layer adapter module ($512 \times 512 \times 2 \approx 0.5$M, where "M" denotes million), and for PromptFL, it is 1e-3, with the length of the learnable prompt vector on the text side being $16 \times 512 \approx 0.008$M.

For the Fed-Prostate benchmark, all data are pre-processed to standardize the field of view for the prostate region and resized to 384×384 in the axial plane. We follow the settings from FedDG-ELCFS (Liu QD et al., 2021), using the U-Net (Ronneberger et al., 2015) model with a learning rate of 1e-3, the Adam optimizer, a batch size of 5, a local epoch count of $E = 1$, and a total of 100 rounds. For evaluation, the Dice coefficient is used to quantitatively assess segmentation results across the entire object region.

For the Shakespeare benchmark, following the data split from q-FFL (Li T et al., 2020a), we sample 31 speaking roles to train a deep language model for next-character prediction. The model takes an 80-character sequence as input, embeds each character in an eight-dimensional learned space, and outputs a single character after passing through two long short-term memory (LSTM) layers and a densely connected layer. The learning rate is set to 0.8 with the stochastic gradient descent (SGD) optimizer, a batch size of 10, a local epoch count of $E = 1$, and a total of 80 epochs.

4.2 Main results

We present the results of using a fully-parameterized ResNet50 as the backbone in Tables 2 and 3. The federated training results, leveraging the pre-trained CLIP model through the FedCLIP and PromptFL training frameworks, are detailed in Tables 4–7.

Table 2 In-domain personalization result comparison for local models across five DomainBed benchmarks with state-of-the-art methods using the fully supervised ResNet50

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	96.52	85.04	85.51	93.46	89.12	89.93
Ditto	97.69	86.14	**86.88**	<u>95.52</u>	<u>90.88</u>	<u>91.42</u>
AFL	96.56	86.36	<u>86.26</u>	95.21	90.94	91.07
q-FFL	97.50	86.35	85.20	91.88	<u>90.88</u>	90.36
PerFedAvg	97.23	<u>86.87</u>	85.76	94.81	90.27	90.99
GRACE	<u>97.88</u>	86.18	85.79	94.93	89.83	90.92
FFT-GP	**97.98**	**87.12**	<u>86.26</u>	**95.63**	**91.18**	**91.63**

Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 3 Out-of-domain generalization result comparison for the global model across five DomainBed benchmarks with state-of-the-art methods using the fully supervised ResNet50

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	86.16	78.21	70.64	44.55	77.27	71.37
FedProx	85.01	77.29	71.19	45.34	77.95	71.36
AFL	85.95	77.19	70.17	44.53	76.85	70.94
q-FFL	85.57	77.95	**71.82**	44.32	78.03	71.54
PerFedAvg	83.89	77.89	68.95	45.28	78.62	70.93
GA	86.62	79.29	69.98	<u>48.67</u>	<u>80.98</u>	<u>73.11</u>
GRACE	<u>86.75</u>	<u>79.46</u>	68.94	45.96	77.78	71.78
FFT-GP	**87.59**	**80.65**	<u>71.43</u>	**50.63**	**81.82**	**74.42**

Bold values indicate the maximum values, while underlined values represent the second-highest values

4.2.1 Overall results

It is apparent that FFT-GP consistently enhances performance in both local model personalization and global model generalization. Specifically, in the realm of in-domain personalization, FFT-GP surpasses methods such as PerFedAvg and Ditto, which primarily focus on local personalization. Furthermore, FFT-GP, by adopting FAA, initiates personalization from a global model with broader applicability. This aligns with the GPFL principle that personalization and generalization can reciprocally reinforce one another. For out-of-domain generalization, FFT-GP, compared with GA, determines aggregation weights with a stronger emphasis on fairness, thus substantially improving the global model's generalizability. In contrast, the global aggregation strategy of GRACE, which prioritizes consistent convergence amid domain shifts without considering fairness, results in poorer performance than both our approach and GA.

Table 4 In-domain personalization result comparison for local models across five DomainBed benchmarks with state-of-the-art methods using the FedCLIP training framework with a ViT-B/16 backbone

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedCLIP	97.72	86.81	84.82	70.92	91.62	86.38
Ditto	**98.55**	**90.15**	88.43	73.51	**92.76**	**88.68**
PerFedAvg	98.27	89.49	88.53	**74.07**	91.68	88.41
GRACE	98.27	89.45	88.51	73.42	91.75	88.28
FFT-GP	98.37	89.59	**88.76**	74.03	91.95	88.54

Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 5 Out-of-domain generalization comparison of the global model across five DomainBed benchmarks with state-of-the-art methods using the FedCLIP training framework with a ViT-B/16 backbone

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedCLIP	97.38	83.88	84.81	48.24	85.02	79.87
FedProx	97.21	83.51	84.86	50.49	84.68	80.15
PerFedAvg	97.38	83.97	85.03	50.80	84.51	80.34
GA	97.23	84.42	85.14	51.45	86.03	80.85
GRACE	97.38	84.28	84.77	50.88	85.19	80.50
FFT-GP	**97.50**	**84.75**	**86.17**	**51.81**	**86.70**	**81.39**
Zero-shot	96.15	81.74	82.19	33.41	88.18	76.33

Bold values indicate the maximum values, while underlined values represent the second-highest values

4.2.2 Supervised vs. CLIP-based

When comparing the fully-parameterized training detailed in Tables 2 and 3 with the pre-trained CLIP-based training presented in Tables 4–7, we notice significant differences. Specifically, for datasets such as PACS, VLCS, OfficeHome, and DomainNet, which are sourced from the Internet, aligning the training data distribution with the pre-trained large model enables impressive personalization and generalization results. This is evident even when fine-tuning a minimal number of parameters (0.5M for FedCLIP and 0.008M for PromptFL). Conversely, for TerraInc, characterized by images from outdoor cameras across diverse locations, a substantial domain shift is observed. Here, using fewer fine-tuning parameters leads to a reduced personalization performance compared with the full-parameter training with ResNet50. However, the robust generalization capability inherent in CLIP-based models allows their global generalization to outperform that of ResNet50.

Table 6 In-domain personalization result comparison for local models across five DomainBed benchmarks with state-of-the-art methods using the PromptFL training framework with a ViT-B/16 backbone

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
PromptFL	98.10	88.63	87.98	**71.56**	90.98	87.45
Ditto	98.13	90.09	88.34	70.32	92.22	87.82
PerFedAvg	98.52	89.06	88.26	70.19	**92.49**	87.70
GRACE	98.66	89.92	**88.88**	70.79	92.26	88.10
FFT-GP	**98.68**	**90.27**	88.82	71.10	92.32	**88.24**

Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 7 Out-of-domain generalization comparison of the global model across five DomainBed benchmarks with state-of-the-art methods using the PromptFL training framework with a ViT-B/16 backbone

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
PromptFL	97.40	82.03	83.71	48.52	85.86	79.50
FedProx	97.69	82.27	83.74	47.28	85.02	79.20
PerFedAvg	97.44	82.39	83.24	44.64	86.03	78.75
GA	97.61	83.49	84.24	50.90	**88.38**	80.92
GRACE	98.03	84.55	85.25	47.16	85.86	80.17
FFT-GP	**98.17**	**86.43**	**86.42**	**51.16**	**88.38**	**82.11**
Zero-shot	96.15	81.74	82.19	33.41	88.18	76.33

Bold values indicate the maximum values, while underlined values represent the second-highest values

4.2.3 FedCLIP vs. PromptFL

Comparing the results of two distinct federated training frameworks using CLIP, we find that FedCLIP outperforms PromptFL in terms of personalization performance with its larger number of fine-tuning parameters (see Tables 4 and 6). However, FedCLIP is more susceptible to overfitting the training data, leading to only marginal improvements on the out-of-domain data compared to PromptFL which uses fewer parameters (see Tables 5 and 7). Additionally, we showcase the averaged zero-shot testing results employing the original CLIP model, illustrating that federated learning can markedly boost the model's generalization ability across different domains. Further insights are offered in the ablation study depicted in Fig. 4, which clarifies the relationship between the number of fine-tuning parameters and the effectiveness of federated training.

4.2.4 Results on more modalities

Tables 8–10 present the experimental results for Fed-Prostate and Shakespeare. On the Fed-Prostate benchmark, methods employing a meta-learning

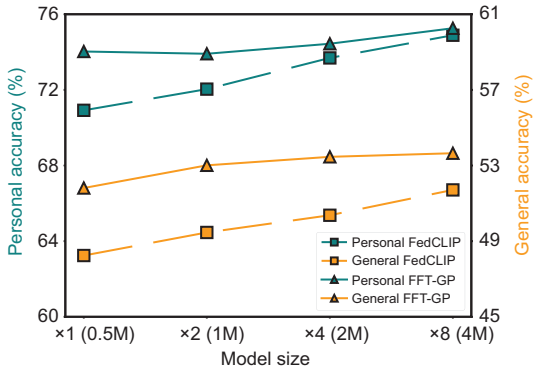

Fig. 4 Results on different FedCLIP fine-tuning parameter sizes: green for personalization (left axis) and yellow for generalization (right axis). "×1 (0.5M)" is the standard size, where "M" indicates million. CLIP: contrastive language-image pre-training. References to color refer to the online version of this figure

Table 8 In-domain personalization result comparison for local models with state-of-the-art methods for the Fed-Prostate benchmark

Method	Dice (%)						
	Domain=*A*	*B*	*C*	*D*	*E*	*F*	Avg.
FedAvg	90.71	90.29	89.65	91.36	89.50	89.30	90.14
Ditto	91.09	90.83	90.91	91.55	91.23	89.87	90.91
AFL	86.38	87.46	85.92	89.15	86.02	86.02	86.83
q-FFL	88.39	85.81	87.18	90.20	87.23	86.21	87.50
PerFedAvg	93.54	**93.93**	93.75	94.22	93.89	93.40	93.79
GRACE	**93.56**	93.72	**94.08**	**94.23**	**93.91**	**93.44**	**93.82**
FFT-GP	93.55	93.74	93.67	94.04	93.64	93.37	93.67

Bold values indicate the maximum values, while underlined values represent the second-highest values

strategy (PerFedAvg, GRACE, and our FFT-GP) achieve the highest levels of in-domain personalization. However, due to the differences in global aggregation and local training strategies, FFT-GP delivers the best out-of-domain generalization with the global model. On the Shakespeare benchmark, FFT-GP also achieves the best average result. Its fairness-guided aggregation strategy significantly reduces the performance variance across different clients, leading to a marked improvement in the "worst 10%" of clients.

4.3 Ablation studies

In this subsection, we describe our analysis of the impact of the FFT-GP algorithm's components on personalization and generalization. We then present the results concerning the reduction of the overall communication cost by alternating between GA and FAA for aggregation. Additionally, we explore the effects of using various sizes of fine-tuning parameters on the personalization and generalization performances in FedCLIP. Finally, we conduct a visual comparison to assess the generalization capabilities of the global model, including an analysis of the distribution differences between training and test data, as well as the distribution variances of the global model on the loss function surface in the test domain.

Table 9 Out-of-domain generalization comparison of the global model with state-of-the-art methods for the Fed-Prostate benchmark

Method	Dice (%)						
	Domain=*A*	*B*	*C*	*D*	*E*	*F*	Avg.
FedAvg	89.57	88.63	82.69	85.39	79.21	89.74	85.87
FedProx	90.55	87.69	83.27	85.42	79.05	90.18	86.03
AFL	82.91	85.19	80.83	81.77	80.69	87.39	83.13
q-FFL	84.30	82.64	81.01	84.33	80.73	87.96	83.50
PerFedAvg	92.08	89.63	84.94	87.11	78.17	89.35	86.88
GA	91.82	89.63	85.23	87.81	81.19	90.23	87.65
GRACE	91.53	90.15	84.28	87.55	81.39	**92.37**	87.88
FFT-GP	**92.43**	**90.77**	**87.55**	**88.49**	**83.09**	91.18	**88.92**

Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 10 Accuracy comparison with state-of-the-art methods for the Shakespeare benchmark

Method	Average	Worst 10%	Best 10%	Variance
FedAvg	46.3	36.5	**71.7**	74
Ditto	50.9	41.5	69.3	46
AFL	35.8	23.5	57.3	58
q-FFL	52.1	42.1	69.0	54
PerFedAvg	47.9	39.0	**71.7**	64
GA	52.1	42.9	69.5	48
GRACE	48.1	39.1	67.5	49
FFT-GP	**53.5**	**46.4**	69.3	**35**

Bold values indicate the maximum values, while underlined values represent the second-highest values

4.3.1 Impact of BGP and FAA

As illustrated in Tables 11 and 12, solely employing FAA aggregation enhances the out-of-domain generalization performance of the global model, which in turn benefits the training of local models. As detailed in the ablation study specific to the BGP framework in Tables 13 and 14, employing only meta-learning ("only ML") yields a notable improvement in in-domain personalization performance. However, this approach tends to compromise the generalization of the global model. Conversely, using only domain alignment ("only DA") results in an in-domain performance similar to that of FedAvg, yet it enhances out-of-domain generalization due to the alignment of feature distributions during local training. The exclusive use of BGP, which combines ML and DA during the local training phase, significantly enhances the personalization performance and

also contributes to the global model's generalization. Finally, the combination of FAA for global aggregation and BGP for local training is complementary, and yields the best results.

Table 11 Ablation study of in-domain personalization results on different components of FFT-GP, encompassing BGP for local models and FAA for the global model, using supervised ResNet50 across five benchmarks

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	96.52	85.04	85.51	93.46	89.12	89.93
BGP	97.93	86.48	86.00	95.08	90.34	91.17
FAA	96.82	85.58	85.40	95.20	89.93	90.59
FFT-GP	**97.98**	**87.12**	**86.26**	**95.63**	**91.18**	**91.63**

Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 12 Ablation study of out-of-domain generalization on different components of FFT-GP, encompassing BGP for local models and FAA for the global model, using supervised ResNet50 across five benchmarks

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	86.16	78.21	70.64	44.55	77.27	71.37
BGP	86.65	79.54	70.44	46.09	79.97	72.54
FAA	87.23	80.53	70.48	49.78	81.32	73.87
FFT-GP	**87.59**	**80.65**	**71.43**	**50.63**	**81.82**	**74.42**

Bold values indicate the maximum values, while underlined values represent the second-highest values

4.3.2 Different levels of communication cost

Table 15 illustrates the effects of alternating between GA and FAA aggregation strategies on the generalizability of the TerraInc dataset. In the table, the fractions 1/2, 1/4, 1/8, and 1/20 indicate that FAA is used in the last round of every 2, 4, 8, and 20 rounds, respectively, with GA being used in the remaining rounds. Given that GA and FedAvg incur identical communication overhead, this approach effectively minimizes additional costs. The results clearly show that an increase in the frequency of FAA usage correlates with enhanced generalization performance of the global model. This strategy offers a pragmatic method to balance performance and efficiency, allowing for optimization based on cost constraints in real-world applications.

In Fig. 4, we illustrate how, within the FedCLIP framework, increasing the width of the two-layer multi-layer perceptron (MLP) network in the extended Adapter results in a corresponding increase in the number of fine-tuning parameters, with widths expanded by factors of 2, 4, and 8, raising the number of fine-tuning parameters to 1M, 2M, and 4M, respectively. It is observed that both personalization and generalization performances enhance as the number of trainable parameters increases. When the number of parameters increases, the disparity in the personalization performance between the two algorithms diminishes; however, a notable difference in the generalization performance of the global model persists.

Table 13 Ablation study of in-domain personalization results for local models of BGP with the fully supervised ResNet-50 across five DomainBed benchmarks

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	96.52	85.04	85.51	93.46	89.12	89.93
Only ML	97.23	86.87	85.76	94.81	90.27	90.99
Only DA	95.96	86.37	85.54	92.73	88.72	89.86
BGP	97.93	86.48	86.00	95.08	90.34	91.17
FFT-GP	**97.98**	**87.12**	**86.26**	**95.63**	**91.18**	**91.63**

Only ML denotes using only meta-learning, whereas only DA indicates using only domain alignment. Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 14 Ablation study of out-of-domain generalization results for the global model of BGP with the fully supervised ResNet-50 across five DomainBed benchmarks

Method	Accuracy (%)					
	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
FedAvg	86.16	78.21	70.64	44.55	77.27	71.37
Only ML	83.89	77.89	68.95	45.28	78.62	70.93
Only DA	85.33	78.11	69.50	46.67	78.79	71.68
BGP	86.65	79.54	70.44	46.09	79.97	72.54
FFT-GP	**87.59**	**80.65**	**71.43**	**50.63**	**81.82**	**74.42**

Only ML denotes using only meta-learning, whereas only DA indicates using only domain alignment. Bold values indicate the maximum values, while underlined values represent the second-highest values

Table 15 The generalization results of using GA to varying extents to replace FAA during the aggregation phase

Method	Extra cost	Accuracy (%)				
		L100	L38	L43	L46	Avg.
GA	0	62.59	45.53	46.73	39.85	48.68
FAA $\frac{1}{20}$	+5%	61.54	42.92	49.50	43.16	49.28
FAA $\frac{1}{8}$	+12.5%	61.85	43.99	50.38	42.82	49.76
FAA $\frac{1}{4}$	+25%	61.96	43.63	49.24	**44.94**	49.94
FAA $\frac{1}{2}$	+50%	62.80	**45.79**	51.64	43.84	**51.02**
FAA	+100%	**63.33**	44.05	**53.27**	41.89	50.64

Experiments were conducted on the TerraInc benchmark using ResNet50. Bold values indicate the maximum values, while underlined values represent the second-highest values

4.3.3 Visualization of loss landscapes on the unseen clients

We use the dimensionality reduction visualization technique (van der Maaten and Hinton, 2008), akin to the one used in GA, to depict the loss landscapes on the unseen client in Fig. 5. It illustrates the distribution of global and local models obtained with three methods: FedAvg, GA, and our FFT-GP. It is apparent that by considering fairness among training nodes during the aggregation process, the global models of GA and FAA tend to converge towards flatter areas of the loss function surface, resulting in enhanced generalization performance.

4.3.4 Curves of fairness

We evaluate the fairness in the performance distribution of the global model under various aggregation methods, as depicted in Fig. 6. The variance of the generalization gap $\{G_i\}_{i=1}^{M}$ is used to depict the performance distribution fairness. It is evident that FedAvg, with its constant weight aggregation, leads to a higher variance in the performance of the global model across different clients, signifying a reduced fairness. Although GA demonstrates an improvement over FedAvg, FAA exhibits a more substantial reduction in variance, with the variance nearing zero in most instances.

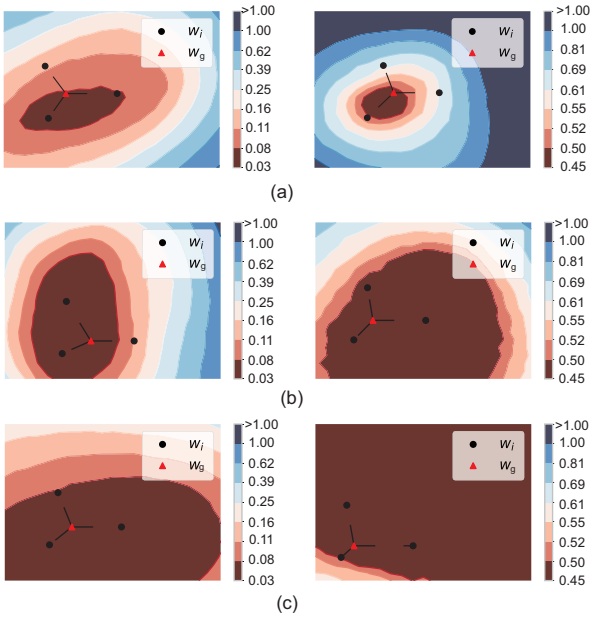

Fig. 5 Visualization of loss function surfaces for various algorithms, FedAvg (a), GA (b), and our FFT-GP (c), with all experiments performed on the PACS benchmark using ResNet50 as the backbone. The test domain on the left is "photo," while the test domain on the right is "art painting." FFT-GP: fairness-guided federated training for generalization and personalization; GA: generalization adjustment

4.3.5 t-SNE visualization of in-domain and out-of-domain features

We conduct a t-SNE dimensionality (van der Maaten and Hinton, 2008) reduction visualization to explore the feature distribution of global models derived from various aggregation algorithms on in-domain training data and out-of-domain test data. It is clear in Fig. 7 that FedAvg, which overlooks the domain shift issue, displays a marked disparity in the feature distributions of training and test data. Conversely, GA and FAA mitigate this disparity, as evidenced by the greater convergence of feature distributions between the training and test data.

5 Conclusions

In this paper, we assert for the first time that it is possible to achieve the GPFL objective under both traditional and CLIP-based federated training, simultaneously enhancing the cross-domain generalization of the global model and the personalization performance of local training clients. We explore this from the perspective of enhancing the performance distribution fairness of the federated system, conducting further observations on the relationship between the generalization gap and aggregation weights as proposed in the previous work, GA (Zhang RP et al., 2023a), and discovering an explicit linear relationship. Based on these observations, we propose the FFT-GP method. Our approach employs an FAA method that minimizes the variance of the generalization gap on training clients and introduces a meta-learning strategy that incorporates constraints for aligning with the global model's feature distribution during local training, thus achieving a balance between generalization and personalization. Through extensive experiments, FFT-GP demonstrates superior performance over existing methods, showcasing its potential to enhance FL systems in various real-world applications.

Contributors

Ruipeng ZHANG designed the research. Ruipeng ZHANG and Ziqing FAN conducted the experiments.

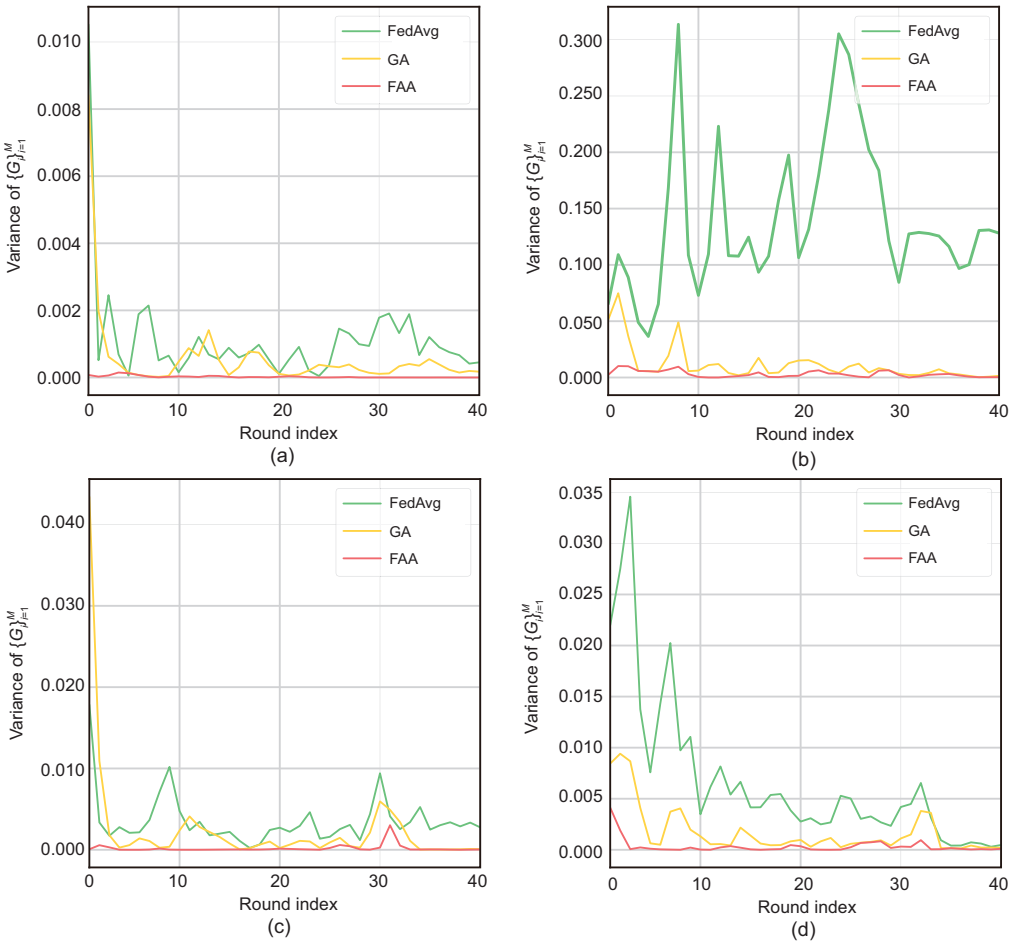

Fig. 6 Fairness curves of global models, obtained using different aggregation methods (FedAvg, GA, and FAA) during local training: (a) PACS; (b) VLCS; (c) OfficeHome; (d) TerraInc. Notably, FFT-GP is the foundation of the training process. After each round of local training, we evaluate fairness through different aggregation algorithms, with the results from FAA guiding the aggregation strategy for the subsequent experimental round. FAA: fairness-aware aggregation; FFT-GP: fairness-guided federated training for generalization and personalization; GA: generalization adjustment. References to color refer to the online version of this figure

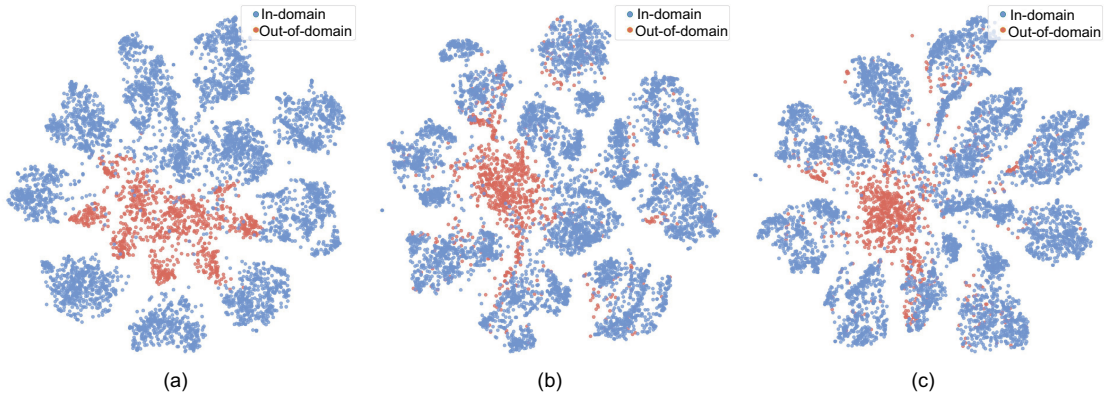

Fig. 7 Visualization of the global model's feature dimensionality reduction using t-SNE under various aggregation methods: (a) FedAvg; (b) GA; (c) FAA. Blue denotes in-domain training data and red represents unseen out-of-domain test data. The visualization is based on the DomainNet benchmark with a ResNet50 backbone, with "quickdraw" as the test domain. References to color refer to the online version of this figure

Ruipeng ZHANG drafted the paper. Ziqing FAN and Jiangchao YAO helped revise the paper. Ya ZHANG and Yanfeng WANG supervised the project and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in several public repositories: DomainBed at https://github.com/facebookresearch/DomainBed, LEAF at https://github.com/TalwalkarLab/leaf, and SAML at https://liuquande.github.io/SAML/.

References

Achiam J, Adler S, Agarwal S, et al., 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774

Arivazhagan MG, Aggarwal V, Singh AK, et al., 2019. Federated learning with personalization layers. https://arxiv.org/abs/1912.00818

Beery S, Van Horn G, Perona P, 2018. Recognition in terra incognita. Proc 15th European Conf on Computer Vision, p.456-473. https://doi.org/10.1007/978-3-030-01270-0_28

Bloch N, Madabhushi A, Huisman H, et al., 2015. NCI-ISBI 2013 challenge: automated segmentation of prostate structures. https://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv

Caldas S, Duddu SMK, Wu P, et al., 2018. LEAF: a benchmark for federated settings. https://arxiv.org/abs/1812.01097

Chu LY, Wang LJ, Dong YJ, et al., 2021. FedFair: training fair models in cross-silo federated learning. https://arxiv.org/abs/2109.05662

Cohen JP, Hashir M, Brooks R, et al., 2020. On the limits of cross-domain generalization in automated X-ray prediction. Proc Int Conf on Medical Imaging with Deep Learning, p.136-155.

Collins L, Hassani H, Mokhtari A, et al., 2021. Exploiting shared representations for personalized federated learning. Proc 38th Int Conf on Machine Learning, p.2089-2099.

Cong Y, Qiu J, Zhang K, et al., 2023. Ada-FFL: adaptive computing fairness federated learning. *CAAI Trans Intell Technol*, 9(3):541-584. https://doi.org/10.1049/cit2.12232

Cui S, Pan WS, Liang J, et al., 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. Proc 35th Int Conf on Neural Information Processing Systems, p.26091-26102.

Ding N, Qin YJ, Yang G, et al., 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell*, 5(3):220-235. https://doi.org/10.1038/s42256-023-00626-4

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. Proc 9th Int Conf on Learning Representations, p.1-22.

du Terrail JO, Ayed SS, Cyffers E, et al., 2022. FLamby: datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. Proc 36th Int Conf on Neural Information Processing Systems, p.1-20.

Fallah A, Mokhtari A, Ozdaglar AE, 2020. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. Proc 34th Int Conf on Neural Information Processing Systems.

Fan ZQ, Wang YF, Yao JC, et al., 2022. FedSkip: combatting statistical heterogeneity with federated skip aggregation. Proc IEEE Int Conf on Data Mining, p.131-140. https://doi.org/10.1109/ICDM54844.2022.00023

Fan ZQ, Zhang RP, Yao JC, et al., 2023a. Federated learning with bilateral curation for partially class-disjoint data. Proc 37th Int Conf on Neural Information Processing Systems.

Fan ZQ, Yao JC, Zhang RP, et al., 2023b. Federated learning under partially class-disjoint data via manifold reshaping. Proc Transactions on Machine Learning Research.

Fang C, Xu Y, Rockmore DN, 2013. Unbiased metric learning: on the utilization of multiple datasets and web images for softening bias. Proc IEEE Int Conf on Computer Vision, p.1657-1664. https://doi.org/10.1109/ICCV.2013.208

Gulrajani I, Lopez-Paz D, 2021. In search of lost domain generalization. Proc 9th Int Conf on Learning Representations, p.1-27.

Guo T, Guo S, Wang JX, et al., 2024. PromptFL: let federated participants cooperatively learn prompts instead of models—federated learning in age of foundation model. *IEEE Trans Mob Comput*, 23(5):5179-5194. https://doi.org/10.1109/TMC.2023.3302410

Haque A, Milstein A, Fei-Fei L, 2020. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193-202. https://doi.org/10.1038/s41586-020-2669-y

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Huang C, Huang JW, Liu X, 2022. Cross-silo federated learning: challenges and opportunities. https://arxiv.org/abs/2206.12949

Huang YT, Chu LY, Zhou ZR, et al., 2021. Personalized cross-silo federated learning on non-IID data. Proc 35th AAAI Conf on Artificial Intelligence, p.7865-7873. https://doi.org/10.1609/aaai.v35i9.16960

Jiang MR, Yang HZ, Cheng C, et al., 2023. IOP-FL: inside-outside personalization for federated medical image segmentation. *IEEE Trans Med Imag*, 42(7):2106-2117. https://doi.org/10.1109/TMI.2023.3263072

Kairouz P, McMahan HB, Avent B, et al., 2021. Advances and open problems in federated learning. *Found Trends Mach Learn*, 14(1-2):1-210. https://doi.org/10.1561/2200000083

Karimireddy SP, Kale S, Mohri M, et al., 2020. SCAFFOLD: stochastic controlled averaging for federated learning. Proc 37th Int Conf on Machine Learning, p.5132-5143.

Khosla A, Zhou TH, Malisiewicz T, et al., 2012. Undoing the damage of dataset bias. Proc 12th European Conf on Computer Vision, p.158-171. https://doi.org/10.1007/978-3-642-33718-5_12

Kirillov A, Mintun E, Ravi N, et al., 2023. Segment anything. https://arxiv.org/abs/2304.02643

Lemaître G, Martí R, Freixenet J, et al., 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput Biol Med*, 60:8-31. https://doi.org/10.1016/j.compbiomed.2015.02.009

Li D, Yang YX, Song YZ, et al., 2017. Deeper, broader and artier domain generalization. Proc IEEE Int Conf on Computer Vision, p.5542-5550. https://doi.org/10.1109/ICCV.2017.591

Li T, Sanjabi M, Beirami A, et al., 2020a. Fair resource allocation in federated learning. Proc 8th Int Conf on Learning Representations, p.1-27.

Li T, Sahu AK, Talwalkar A, et al., 2020b. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*, 37(3):50-60. https://doi.org/10.1109/MSP.2020.2975749

Li T, Sahu AK, Zaheer M, et al., 2020c. Federated optimization in heterogeneous networks. https://arxiv.org/abs/1812.06127v5

Li T, Hu SY, Beirami A, et al., 2021. Ditto: fair and robust federated learning through personalization. Proc 38th Int Conf on Machine Learning, p.6357-6368.

Li X, Huang KX, Yang WH, et al., 2020. On the convergence of FedAvg on non-IID data. Proc 8th Int Conf on Learning Representations, p.1-26.

Li XX, Jiang MR, Zhang XF, et al., 2021. FedBN: federated learning on non-IID features via local batch normalization. Proc 9th Int Conf on Learning Representations, p.1-27.

Litjens G, Toth R, van de Ven W, et al., 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med Image Anal*, 18(2):359-373. https://doi.org/10.1016/j.media.2013.12.002

Liu KZ, Hu SY, Wu S, et al., 2022. On privacy and personalization in cross-silo federated learning. Proc 36th Int Conf on Neural Information Processing Systems, p.5925-5940.

Liu QD, Dou Q, Yu LQ, et al., 2020. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE Trans Med Imag*, 39(9):2713-2724. https://doi.org/10.1109/TMI.2020.2974574

Liu QD, Chen C, Qin J, et al., 2021. FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1013-1023. https://doi.org/10.1109/CVPR46437.2021.00107

Lu W, Hu XX, Wang JD, et al., 2023. FedCLIP: fast generalization and personalization for CLIP in federated learning. Proc Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models, p.1-14.

Lyu HQ, Zhang YX, Wang C, et al., 2023. Federated learning privacy incentives: reverse auctions and negotiations. *CAAI Trans Intell Technol*, 8(4):1538-1557. https://doi.org/10.1049/cit2.12190

Ma ZZ, Zhao MY, Cai XJ, et al., 2021. Fast-convergent federated learning with class-weighted aggregation. *J Syst Archit*, 117:102125. https://doi.org/10.1016/j.sysarc.2021.102125

McMahan B, Moore E, Ramage D, et al., 2017. Communication-efficient learning of deep networks from decentralized data. Proc 20th Int Conf on Artificial Intelligence and Statistics, p.1273-1282.

Mohri M, Sivek G, Suresh AT, 2019. Agnostic federated learning. Proc 36th Int Conf on Machine Learning, p.4615-4625.

Nguyen AT, Torr PHS, Lim SN, 2022. FedSR: a simple and effective domain generalization method for federated learning. Proc 36th Int Conf on Neural Information Processing Systems, p.38831-38843.

Oh J, Kim S, Yun SY, 2022. FedBABU: toward enhanced representation for federated image classification. Proc 10th Int Conf on Learning Representations, p.1-29.

Peng XC, Bai QX, Xia XD, et al., 2019. Moment matching for multi-source domain adaptation. Proc IEEE/CVF Int Conf on Computer Vision, p.1406-1415. https://doi.org/10.1109/ICCV.2019.00149

Radford A, Kim JW, Hallacy C, et al., 2021. Learning transferable visual models from natural language supervision. Proc 38th Int Conf on Machine Learning, p.8748-8763.

Rieke N, Hancox J, Li WQ, et al., 2020. The future of digital health with federated learning. *npj Digit Med*, 3(1):119. https://doi.org/10.1038/s41746-020-00323-1

Ronneberger O, Fischer P, Brox T, 2015. U-Net: convolutional networks for biomedical image segmentation. Proc 18th Medical Image Computing and Computer-Assisted Intervention, p.234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Schuhmann C, Beaumont R, Vencu R, et al., 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. Proc 36th Int Conf on Neural Information Processing Systems, p.25278-25294.

Shi YX, Yu H, Leung C, 2024. Towards fairness-aware federated learning. *IEEE Trans Neur Netw Learn Syst*, 35(9):11922-11938. https://doi.org/10.1109/TNNLS.2023.3263594

Smith V, Chiang CK, Sanjabi M, et al., 2017. Federated multi-task learning. Proc 30th Int Conf on Neural Information Processing Systems.

Su SC, Yang MZ, Li B, et al., 2024. Federated adaptive prompt tuning for multi-domain collaborative learning. Proc 38th AAAI Conf on Artificial Intelligence, 38:15117-15125. https://doi.org/10.1609/aaai.v38i13.29434

Sun BC, Saenko K, 2016. Deep CORAL: correlation alignment for deep domain adaptation. European Conf on Computer Vision, p.443-450. https://doi.org/10.1007/978-3-319-49409-8_35

van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. *J Mach Learn Res*, 9(86):2579-2605.

Venkateswara H, Eusebio J, Chakraborty S, et al., 2017. Deep hashing network for unsupervised domain adaptation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5018-5027. https://doi.org/10.1109/CVPR.2017.572

Wang JY, Liu QH, Liang H, et al., 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. Proc 34th Int Conf on Neural Information Processing Systems, p.7611-7623.

Wei GYZ, Wang F, Shah A, et al., 2023. Dual prompt tuning for domain-aware federated learning. https://arxiv.org/abs/2310.03103

Xu A, Li WQ, Guo PF, et al., 2022. Closing the generalization gap of cross-silo federated medical image segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.20866-20875. https://doi.org/10.1109/CVPR52688.2022.02020

Xu QW, Zhang RP, Zhang Y, et al., 2021. A Fourier-based framework for domain generalization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.14383-14392. https://doi.org/10.1109/CVPR46437.2021.01415

Xu QW, Zhang RP, Zhang Y, et al., 2024. Federated adversarial domain hallucination for privacy-preserving domain generalization. *IEEE Trans Multim*, 26:1-14. https://doi.org/10.1109/TMM.2023.3257566

Yuan HL, Morningstar WR, Ning L, et al., 2022. What do we mean by generalization in federated learning? Proc 10th Int Conf on Learning Representations, p.1-26.

Yuan JK, Ma X, Chen DF, et al., 2023. Collaborative semantic aggregation and calibration for federated domain generalization. *IEEE Trans Knowl Data Eng*, 35(12):12528-12541. https://doi.org/10.1109/TKDE.2023.3271851

Zeng YC, Chen HX, Lee K, 2021. Improving fairness via federated learning. https://arxiv.org/abs/2110.15545v2

Zhang FD, Kuang K, Chen L, et al., 2023. Federated unsupervised representation learning. *Front Inform Technol Electron Eng*, 24(8):1181-1193. https://doi.org/10.1631/FITEE.2200268

Zhang FD, Shuai ZT, Kuang K, et al., 2024. Unified fair federated learning for digital healthcare. *Patterns*, 5(1):100907. https://doi.org/10.1016/J.PATTER.2023.100907

Zhang HR, Dullerud N, Seyyed-Kalantari L, et al., 2021. An empirical framework for domain generalization in clinical settings. Proc Conf on Health, Inference, and Learning, p.279-290. https://doi.org/10.1145/3450439.3451878

Zhang RP, Xu QW, Yao JC, et al., 2023a. Federated domain generalization with generalization adjustment. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.3954-3963. https://doi.org/10.1109/CVPR52729.2023.00385

Zhang RP, Fan ZQ, Xu QW, et al., 2023b. GRACE: a generalized and personalized federated learning method for medical imaging. Proc 26th Int Conf on Medical Image Computing and Computer-Assisted Intervention, p.14-24. https://doi.org/10.1007/978-3-031-43898-1_2

Zhao Y, Li M, Lai LZ, et al., 2018. Federated learning with non-IID data. https://arxiv.org/abs/1806.00582

Zhou KY, Yang JK, Loy CC, et al., 2022. Learning to prompt for vision-language models. *Int J Comput Vis*, 130(9):2337-2348. https://doi.org/10.1007/s11263-022-01653-1

Zhu HY, Xu JJ, Liu SQ, et al., 2021. Federated learning on non-IID data: a survey. *Neurocomputing*, 465:371-390. https://doi.org/10.1016/j.neucom.2021.07.098

Zhuang WM, Chen C, Lyu LJ, 2023. When foundation model meets federated learning: motivations, challenges, and future directions. https://arxiv.org/abs/2306.15546