pFedMxF: Personalized Federated Class-incremental Learning with Mixture of Frequency Aggregation

Yifei Zhang[1,*] Hao Zhu[2,*] Alysa Ziying Tan[1] Dianzhi Yu[3] Longtao Huang[4] Han Yu[1,†]
[1]College of Computing and Data Science, Nanyang Technological University
[2]Data61♥CSRIO [3]The Chinese University of Hong Kong [4]Alibaba Group
{yifei.zhang, han.yu}@ntu.edu.sg

Abstract

Federated learning (FL) has emerged as a promising paradigm for privacy-preserving collaborative machine learning. However, extending FL to class incremental learning settings introduces three key challenges: 1) spatial heterogeneity due to non-IID data distributions across clients, 2) temporal heterogeneity due to sequential arrival of tasks, and 3) resource heterogeneity due to diverse client capabilities. Existing approaches generally address these challenges in isolation, potentially leading to interference between updates, catastrophic forgetting or excessive communication overhead. In this paper, we propose personalized Federated class-incremental parameter efficient fine-tuning with Mixture of Frequency aggregation (pFedMxF), a novel framework that simultaneously addresses all three heterogeneity challenges through frequency domain decomposition. Our key insight is that assigning orthogonal frequency components to different clients and tasks enables interference-free learning to be achieved with minimal communication costs. We further design an Auto-task Agnostic Classifier that automatically routes samples to task-specific classifiers while adapting to heterogeneous class distributions. We conduct extensive experiments on three benchmark datasets, comparing our approach with eight state-of-the-art methods. The results demonstrate that pFedMxF achieves comparable test accuracy, while requiring less model parameters and incurring significantly lower communication costs than baseline methods.

1. Introduction

Federated learning (FL) [10, 32, 51, 55] has emerged as a useful paradigm for collaborative model training while preserving data privacy. In recent years, with increasing emphasis on data protection regulations [43], FL has been widely adopted across various domains, from healthcare to autonomous driving [19, 33]. However, existing FL methods generally operate under a *closed-world assumption* [12], where the number of classes remains fixed throughout training. This assumption might not always hold in real-world scenarios, where data owners (i.e., FL clients) continuously encounter new classes and update their models accordingly.

The intersection of federated and incremental learning introduces unique challenges [6]. Local clients not only have different data distributions (non-IID data), but might also encounter new classes at different times. Moreover, clients vary in their computational resources and communication capabilities. These variations create a complex landscape of heterogeneity. Firstly, each client's unique data distribution leads to biased local updates that can affect global model performance [62]. Secondly, as new classes emerge over time, the model must learn them without forgetting previously acquired knowledge [31]. Thirdly, varying computational power and communication bandwidth across clients necessitate efficient learning approaches [26].

Current approaches struggle to address these challenges effectively. Some methods store old data samples [6], which compromises privacy and consumes memory. Others generate synthetic data [40–42, 59], which is computationally expensive and often unreliable. While parameter-efficient approaches [16] reduce communication costs, they often suffer from interference across different clients and tasks. The main issue is that adjustments made for one client or task can often disrupt the learning of others.

We draw inspiration from an everyday phenomenon – just as a radio tunes into different stations without interference, what if we assign unique "frequencies" to different clients and tasks to enable independent learning. By translating model updates into the frequency domain, we offer an elegant solution to simultaneously address all three heterogeneity challenges. The frequency domain naturally handles spatial heterogeneity across clients by assigning each one its own distinct frequency components. Much like radio stations broadcasting in different channels, clients can update their models independently, even when their data distributions dif-

[*] Equal Contribution and † Corresponding Author.

fer significantly. This represents a fundamental advancement from traditional methods, which struggle with averaging potentially conflicting updates. When it comes to temporal heterogeneity as new tasks arrive, we can simply tune into new frequencies while keeping the old ones untouched. Much like adding new radio stations without disrupting existing broadcasts, the model can learn new classes while naturally preserving previously learned knowledge, thereby addressing the problem of catastrophic forgetting. Perhaps most importantly, our approach elegantly handles resource heterogeneity through its inherent efficiency. Each client only needs to work with a small slice of the frequency spectrum, keeping computation and communication costs low. This means even devices with limited resources can effectively participate in the learning process, making our approach practical for real-world federated learning scenarios.

Building on these insights, we propose personalized Federated class-incremental Parameter Efficient Fine-Tuning (PEFT) with Mixture of Frequency aggregation (`pFedMxF`). The key contributions include:

1. Proposing a mathematical framework that decomposes parameter updates into orthogonal frequency components, thereby ensuring interference-free learning across both clients and tasks while maintaining minimal communication overhead.
2. Designing an efficient aggregation scheme that enables perfect reconstruction of client updates through frequency mixing, thereby avoiding the information loss in traditional averaging-based methods.
3. Building an Auto-task Agnostic Classifier (AAC) that automatically routes samples to task-specific classifiers, while adapting to heterogeneous class distributions.

Extensive experiments on three benchmark datasets demonstrate that, compared to eight relevant existing methods, `pFedMxF` achieves state-of-the-art test accuracy and superior robustness across different heterogeneity settings, while maintaining constant memory usage regardless of the number of FL clients involved.

2. Related Work

2.1. Federated Class-Incremental Learning

Interest of Federated Class Incremental Learning (FCIL) has grown in recent years. Dong *et al*. [6] introduced this concept and developed loss functions to mitigate both local and global catastrophic forgetting. While effective, their approach relies on storing data from old classes and using a proxy server, resulting in substantial memory and communication costs. Although LGA [7] built upon this work, it remained rehearsal-based. In rehearsal-free FCIL, researchers have explored using generative models to create synthetic data [59]. However, performance is highly dependent on data quality and requires substantial computation. FedSpace [37] took a different approach, using prototype-based loss to cluster same-class features – similar to our prototype classifier method. Recent work [2, 29] has integrated pre-trained models with FCIL, achieving better performance with lower communication costs. However, these approaches use similarity-based selection, causing memory overhead during inference. Additionally, their reliance on supervised pre-training raises privacy concerns, as downstream task data may overlap with pre-training datasets.

2.2. PEFT for Pre-Trained Model

The rise of large-scale pre-trained models [3, 21, 35] has sparked significant interest in parameter efficient fine-tuning (PEFT) methods for downstream tasks. LoRA [16], Prompt [30], and Adapter [15] have emerged as leading techniques, finding widespread application in both CIL [11, 39, 45, 46] and FL [14, 53, 61]. In the context of FCIL, researchers [2, 29] have explored integrating Prompt and Adapter with pre-trained models. These methods store stage-specific knowledge in Prompt or Adapter module parameters and dynamically select appropriate modules during inference through similarity computations. While this approach effectively addresses catastrophic forgetting with minimal communication overhead, the required similarity computations introduce latency during inference.

2.3. LoRA in Continual Learning

As LoRA [16] gains popularity and its variants emerge [5, 50, 60], numerous works in incremental learning (IL) have proposed to integrate LoRA modules into their architectures. This integration allows models to continually acquire new knowledge, such as classifying more classes (CIL), while reducing training costs. Typically, methods freeze the pre-trained base model and train only the LoRA-related modules. In the IL process, Online-LoRA [48] adds a new LoRA module when the loss is stable. Previous LoRA modules are merged into the pre-trained ViT model to reduce training and memory costs further. O-LoRA [44] restricts gradient updates within an orthogonal subspace to past tasks and similarly fixes old LoRA parameters. To further enhance learning capabilities, researchers explore combining the Mixture of Experts (MoE) framework [18] with LoRA. Instead of training one LoRA module, MoE contains multiple “expert" networks and a gating network to select experts, balancing effective knowledge acquisition with computational efficiency. MoRAL [52] employs LoRA expert modules within MoE architecture to facilitate incremental learning for LLMs while maintaining efficient training. Likewise, MoE-Adapters4CL [56] freezes the pre-trained CLIP model [35] and utilizes LoRA experts in the multimodal IL setting (vision and language), which is more challenging than single-modality IL settings [54]. It contains task-specific routers

and develops a selector to decide the proper router for CIL. However, these methods are not in FL setting.

3. Preliminaries

Federated Class Incremental Learning extends conventional class-incremental learning to Federated Learning (FCIL). Let $\mathcal{T} = \{\mathcal{T}^t\}_{t=1}^{t}$ denote sequence of streaming tasks, where t denotes the task number, and the t-th task $\mathcal{T}^t = (\mathcal{X}^t, \mathcal{Y}^t)$ consists of input samples $\boldsymbol{x}^t \in \mathcal{X}$ and labels $y^t \in \mathcal{Y}^t$. $\mathcal{Y}^t$ represents the label space of the t-th task where labels in different task are disjointed ($\mathcal{Y}^t \cap \mathcal{Y}^{t'} = \varnothing$ if $t \neq t'$).

Given κ local clients $\{\mathcal{C}_\kappa\}_{\kappa=1}^{K}$ and a global central server $\mathcal{S}$, for each task t, client $\mathcal{C}_\kappa$ train on the the local dataset $\mathcal{T}_\kappa^t = (\mathcal{X}_\kappa^t, \mathcal{Y}_\kappa^t) \subset \mathcal{T}^t$ via optimizing the following objective

$$\underset{\Delta \boldsymbol{W}_\kappa^t}{\operatorname{argmin}} \mathcal{L}\left(\boldsymbol{W}^{t-1} + \Delta \boldsymbol{W}_\kappa^t;\ \mathcal{X}_\kappa^t, \mathcal{Y}_\kappa^t\right), \tag{1}$$

where $\boldsymbol{W}^{t-1}$ is the parameters of the global model in previous tasks $t-1$ and $\Delta \boldsymbol{W}_\kappa^t$ is the update of local model. The server aggregates all uploaded parameter updates through weighted averaging:

$$\Delta \bar{\boldsymbol{W}}^t = \sum_{\kappa=1}^{K} \gamma_\kappa^t \Delta \boldsymbol{W}_\kappa^t, \text{ where } \gamma_\kappa^t = \frac{|\mathcal{X}_\kappa^t|}{\sum_{\kappa=1}^{K} |\mathcal{X}_\kappa^t|}, \tag{2}$$

and update the global model after each task t as:

$$\boldsymbol{W}^t = \boldsymbol{W}^{t-1} + \Delta \bar{\boldsymbol{W}}^t. \tag{3}$$

However, FCIL faces a fundamental **Challenge** of Heterogeneity in three key aspects:

C1: Spatial Heterogeneity. Due to Non-IID nature of FL, at any task t, the data distribution varies across clients:

$$\mathcal{P}(\mathcal{T}_\kappa^t) \neq \mathcal{P}(\mathcal{T}_{\kappa'}^t), \quad \forall \kappa, \kappa' \text{ where } \kappa \neq \kappa',$$

where clients may have different class sets: $\mathcal{Y}_\kappa^t \neq \mathcal{Y}_{\kappa'}^t$

C2: Temporal Heterogeneity. For each client κ, the data distribution changes across tasks:

$$\mathcal{P}(\mathcal{T}_\kappa^t) \neq \mathcal{P}(\mathcal{T}_\kappa^{t'}), \quad \forall t, t' \text{ where } t \neq t',$$

where class sets are disjoint: $\mathcal{Y}_\kappa^t \cap \mathcal{Y}_\kappa^{t'} = \emptyset$

C3: Resource Heterogeneity: In federated settings, clients have varying computational capabilities and communication bandwidth. The parameter update $\Delta \boldsymbol{W}_\kappa^t \in \mathbb{R}^{d \times k}$ imposes significant resource demands. With heterogeneous client resources, the system performance is often bottlenecked by the most resource-constrained clients ($\mathcal{O}(dk)$ parameters per update), leading to inefficient training and potential client dropouts.

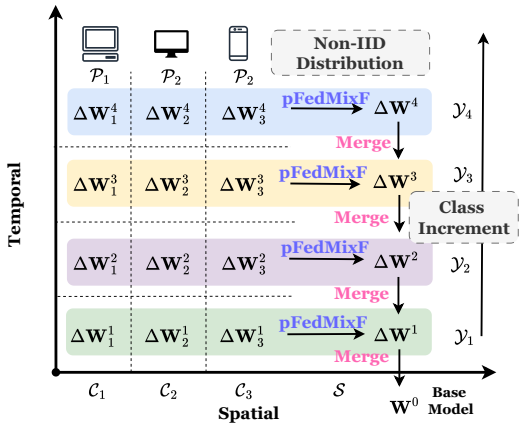


Figure 1. Overview of our pFedMxF framework.

LoRA-based FCIL. To address resource heterogeneity, recent works [16] propose to decompose parameter $\Delta \boldsymbol{W}$ updates through low-rank adaptation[1] [16]:

$$\Delta \bar{\boldsymbol{W}}^t = \sum_{\kappa=1}^{K} \Delta \boldsymbol{W}_\kappa^t, \ \ \Delta \boldsymbol{W}_\kappa^t = \boldsymbol{A}_\kappa^t \boldsymbol{B}_\kappa^t, \tag{4}$$

where $\boldsymbol{A}_\kappa^t \in \mathbb{R}^{d \times r}$, $\boldsymbol{B}_\kappa^t \in \mathbb{R}^{r \times k}$, and $r \ll \min\{d, k\}$. This reduces communication cost from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d+k))$.

Existing aggregation strategies, however, face fundamental limitations:

- **FedAvg** reduces resource requirements through averaging:

$$\Delta \bar{\boldsymbol{W}}_{\text{avg}}^t = \bar{\boldsymbol{A}}^t \bar{\boldsymbol{B}}^t, \text{ where } \bar{\boldsymbol{A}}^t = \sum_{\kappa=1}^{K} \boldsymbol{A}_\kappa^t, \bar{\boldsymbol{B}}^t = \sum_{\kappa=1}^{K} \boldsymbol{B}_\kappa^t. \tag{5}$$

$$\Delta \boldsymbol{W}_{\text{avg}}^t = \Delta \bar{\boldsymbol{W}}^t + \underbrace{\sum_{i=1}^{K} \sum_{j=1}^{K} \boldsymbol{A}_i \boldsymbol{B}_j \ (i \neq j)}_{\text{Interference term}}. \tag{6}$$

  The precise $\Delta \bar{\boldsymbol{W}}$ update will be influenced by the interference term as K increased.

- **FedStack** preserves client-specific updates through stacking:

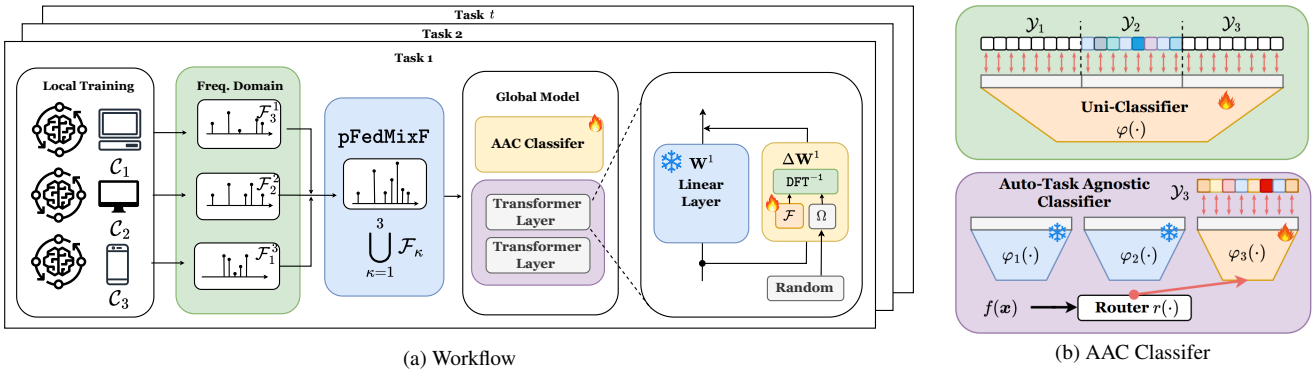
$$\bar{\boldsymbol{A}}^t = [\boldsymbol{A}_1^t, \ldots, \boldsymbol{A}_\kappa^t], \ \ \bar{\boldsymbol{B}}^t = [\boldsymbol{B}_1^t, \ldots, \boldsymbol{B}_\kappa^t]. \tag{7}$$

  This achieves exact aggregation:

$$\Delta \bar{\boldsymbol{W}}_{\text{stack}}^t = \bar{\boldsymbol{A}}^t \bar{\boldsymbol{B}}^t = \sum_{\kappa=1}^{K} \boldsymbol{A}_\kappa^t \boldsymbol{B}_\kappa^t, \tag{8}$$

  but at the cost of increased communication overhead that scales with client number.

[1] we omit the γ_κ^t for simplicity.



(a) Workflow (b) AAC Classifer

Figure 2. Overview of the proposed pFedMxF framework across multiple tasks. The architecture consists of three main components: (1) Local training on heterogeneous clients ($\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$) where each client learns unique frequency components ($\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$) in the frequency domain through 2D-DFT; (2) pFedMxF aggregation that combines orthogonal frequency components from all clients; and (3) Global model update process involving a transformer-based feature extractor, AAC classifier for new classes, and a weight update mechanism that combines the frozen linear layer ($\boldsymbol{W}$) with learned frequency components ($\Delta\boldsymbol{W}$) through inverse 2D DFT.

These heterogeneities present a challenging landscape that existing approaches fail to address effectively. While prior works have tackled these challenges in isolation, there remains a critical need for a unified approach that can simultaneously handle spatial-temporal heterogeneity while operating within the practical resource constraints of FL.

4. Methodology

Our key insight is that heterogeneity challenges can be understood through interference analysis. Consider the loss for different clients/tasks: $\mathcal{L}_\kappa^t = \ell((\mathbf{W}_0 + \Delta\mathbf{W}_\kappa^t)\mathbf{x}_\kappa^t, \mathbf{y}_\kappa^t)$. When data comes from heterogeneous distributions ($\mathbf{x}_\kappa^t \sim \mathcal{P}_\kappa^t$), gradient updates interfere in two dimensions:

$$\begin{aligned} \textbf{Spatial interference:} \quad & \langle \nabla_{\Delta\mathbf{W}} \mathcal{L}_\kappa^t, \nabla_{\Delta\mathbf{W}} \mathcal{L}_{\kappa'}^t \rangle \neq 0 \\ \textbf{Temporal interference:} \quad & \langle \nabla_{\Delta\mathbf{W}} \mathcal{L}_\kappa^t, \nabla_{\Delta\mathbf{W}} \mathcal{L}_\kappa^{t'} \rangle \neq 0 \end{aligned} \tag{9}$$

To avoid such interference while maintaining efficiency, updates must satisfy dual orthogonality conditions:

$$\begin{aligned} \langle \Delta\mathbf{W}_\kappa^t, \Delta\mathbf{W}_{\kappa'}^t \rangle = 0 \quad & \textbf{(Spatial orthogonality)} \\ \langle \Delta\mathbf{W}_\kappa^t, \Delta\mathbf{W}_\kappa^{t'} \rangle = 0 \quad & \textbf{(Temporal orthogonality)} \end{aligned} \tag{10}$$

Based on this analysis, we propose personalized Federated class-incremental parameter efficient fine-tuning with Mixture of Frequency aggregation (pFedMxF).

4.1. Mixture of Frequency Aggregation via 2D DFT

Natural dual orthogonality via 2D DFT basis. We propose addressing Spatial-temporal Heterogeneity by sampling components without replacement in the 2D discrete frequency domain. For a parameter matrix $\Delta\boldsymbol{W} \in \mathbb{R}^{d\times d}$, the 2D DFT (Discrete Fourier Transform) is defined as:

$$F_{(u,v)} = \sum_{m=0}^{d-1}\sum_{n=0}^{d-1} \Delta\boldsymbol{W}(m,n) e^{-j2\pi(\frac{um}{d}+\frac{vn}{d})}, \tag{11}$$

where $F(u,v)$ are the Fourier coefficients learned via backpropagation. The complete set of frequency coordinates is: $\Omega = \{(u,v) | u \in \{0,\cdots,d-1\}, v \in \{0,\cdots,d-1\}\}$. For each client κ at task t, we random sample distinct frequency coordinates per client-task pair without replacement: $\Omega_\kappa^t \sim (\Omega \setminus \bigcup_{(\kappa',t')<(\kappa,t)} \Omega_{\kappa'}^{t'})$ where $(\kappa',t') < (\kappa,t)$ denotes lexicographic ordering. and the set of trainable Fourier coefficients is denote as $\mathcal{F}_\kappa^t = \{F(u,v) | (u,v) \in \Omega_\kappa^t\}$. Since we sample without replacement: $\Omega_\kappa^t \cap \Omega_{\kappa'}^{t'} = \emptyset, \quad \forall(\kappa',t') \neq (\kappa,t)$.

The DFT basis functions are naturally orthogonal:

$$\langle e^{j2\pi(\frac{\kappa m}{M}+\frac{tn}{N})}, e^{j2\pi(\frac{\kappa' m}{M}+\frac{t' n}{N})} \rangle = 0 \text{ if } (\kappa',t') \neq (\kappa,t),$$

which guaranteed the updates $\Delta\boldsymbol{W}_\kappa^t$ reside in **orthogonal subspace** for any task t and client κ:

$$\langle \Delta\boldsymbol{W}_\kappa^t, \Delta\boldsymbol{W}_{\kappa'}^{t'} \rangle = \mathbf{0}, \quad \forall(\kappa,t) \neq (\kappa',t').$$

Mixture of frequency aggregation (pFedMxF). Global update $\Delta\bar{\boldsymbol{W}}^t$ can be viewed as a mixture of local update $\Delta\boldsymbol{W}^t$ that reside in different frequency.

$$\Delta\boldsymbol{W}^t = \underbrace{\Delta\boldsymbol{W}_1^t}_{\mathcal{F}_1^t} + \underbrace{\Delta\boldsymbol{W}_2^t}_{\mathcal{F}_2^t} \cdots \underbrace{\Delta\boldsymbol{W}_\kappa^t}_{\mathcal{F}_\kappa^t}, \tag{12}$$

Therefore, $\Delta\bar{\boldsymbol{W}}^t$ can be easily recovered by mix of frequency aggregation.

$$\begin{aligned} \Delta\bar{\boldsymbol{W}}_{\texttt{MixF}}^t &= \sum_{\kappa=1}^K \Delta\boldsymbol{W}_\kappa^t = \sum_{\kappa=1}^K \sum_{(u,v)\in\Omega_\kappa^t} F(u,v) e^{j2\pi(\frac{um}{M}+\frac{vn}{N})} \\ &= \sum_{(u,v)\in\bigcup_{\kappa=1}^K \Omega_\kappa^t} \bar{F}(u,v) e^{j2\pi(\frac{um}{M}+\frac{vn}{N})}. \end{aligned} \tag{13}$$

Note that $\bar{F}(u, v)$ is the averaged coefficient if frequency coordinates (u, v) is selected in multiple clients.

Overall, pFedMxF offers several unique advantages:

Natural dual orthogonality. The 2D-DFT provides inherent orthogonal bases in both dimensions, while the orthogonality constraint automatically partitions the parameter space (also see figure 5). This prevents catastrophic forgetting caused by task interference and alleviates performance degradation due to non-IID data distributions.

Adaptation to different devices. pFedMxF does not require the number of frequencies $|\Omega|$ to be aligned across devices[b]. Different devices can vary $|\Omega|$ based on their computational resources while still participating in federated learning.

Perfect reconstruction. The method enables lossless recovery of original updates during aggregation, in contrast to FedAvg.

Efficient implementation. pFedMxF achieves efficiency through two key aspects: (1) Each client only needs to learn Fourier coefficients for its sampled frequencies ($\mathcal{O}(|\Omega|)$ parameters), with fast 2D-FFT computation ($\mathcal{O}(|\Omega| \log |\Omega|)$); (2) The orthogonal frequency decomposition ensures constant memory usage $\mathcal{O}(1)$ regardless of the number of clients K, unlike prior methods like FedStack [47] that require storing $\mathcal{O}(K)$ matrices.

[b]LoRA-based FedAvg usually need the same rank to keep the shape of $\boldsymbol{A}$ and $\boldsymbol{B}$ to be identical so that it can perform the average operation

4.2. Auto-task Agnostic Classifier

In addition to the base model (the feature extractor), a new classifier needs to be trained to adapt to downstream tasks. In class-incremental learning, we need to classify samples into the unified label space $\mathcal{Y} = \bigcup_{t=1}^{T} \mathcal{Y}_t$. A principled approach is to model the joint probability distribution $p(\boldsymbol{x}, y, t)$, which can be decomposed as:

$$p(\boldsymbol{x}, y, t) = p(y|\boldsymbol{x}, t) \cdot p(t|\boldsymbol{x}) \cdot p(\boldsymbol{x}). \tag{14}$$

A naive implementation would use a single unified classifier for all classes, computing $p(y|\boldsymbol{x})$ as:

$$p(y|\boldsymbol{x}) = \frac{e^{\varphi(f(\boldsymbol{x}))_y}}{\sum_{y=1}^{|\mathcal{Y}|} e^{\varphi(f(\boldsymbol{x}))_y}}, \tag{15}$$

where $f(\boldsymbol{x}) \in \mathbb{R}^d$ represents the feature embedding of input $\boldsymbol{x}$. However, this approach faces a critical issue:

Task Interference. At each task t, only classes $\mathcal{Y}_t$ are observable and trainable. When using a shared classifier $\varphi(\cdot)$, updates for current task classes inevitably interfere with the decision boundaries of previous tasks, enhancing catastrophic forgetting. Also, in the federated setting, clients may have varying class distributions, making it challenging to maintain consistent classification boundaries across the federation.

Table 1. Efficiency of different federated aggregation methods. K is the number of clients, r is the LoRA rank, d is input/output dimension, and $|\Omega|$ represents # frequency.

Method	#Train Param. (Memory)	# Communication Param.	Time complexity	Precise Aggregation
FedAvg	$\mathcal{O}(rd)$	$\mathcal{O}(Krd)$	$\mathcal{O}(rd)$	No
FedStack	$\mathcal{O}(Krd)$	$\mathcal{O}(K^2rd)$	$\mathcal{O}(Krd)$	Yes
pFedMxF	$\mathcal{O}(\|\Omega\|)$	$\mathcal{O}(K\|\Omega\|)$	$\mathcal{O}(\|\Omega\| \log \|\Omega\|)$	Yes

To address these challenges, we propose the Auto-task Agnostic Classifier (AAC), which explicitly models both components of the decomposed probability. For task-specific classification, we introduce separate classifiers $\varphi_t(\cdot)$ for each task:

$$p(y|\boldsymbol{x}, t) = \frac{e^{\varphi_t(f(\boldsymbol{x}))_y}}{\sum_{y=1}^{|\mathcal{Y}_t|} e^{\varphi_t(f(\boldsymbol{x}))_y}}. \tag{16}$$

To automatically route samples to appropriate task classifiers, we design a router $r(\cdot)$ that estimates task probability based on feature space similarity:

$$r(\boldsymbol{x})_t = p(t|\boldsymbol{x}) \propto \exp\left(-\frac{\|f(\boldsymbol{x}) - \boldsymbol{\mu}_t\|^2}{2}\right), \tag{17}$$

where $\boldsymbol{\mu}_t$ represents the task prototype computed as the mean of classifier parameters $\boldsymbol{\theta} \in \mathbb{R}^{d \times |\mathcal{Y}_t|}$ along class dimensions. Then the final prediction thus becomes:

$$p(y|\boldsymbol{x}) \propto p(y|\boldsymbol{x}, t) \cdot p(t|\boldsymbol{x}) \propto \varphi_t(\boldsymbol{x})_y \cdot r(\boldsymbol{x})_t. \tag{18}$$

In the federated setting, we aggregate local task-specific classifiers through weighted averaging:

$$\bar{\varphi}_t(\boldsymbol{x}) = \sum_{\kappa=1}^{K} \gamma_\kappa \varphi_t^\kappa(\boldsymbol{x}), \tag{19}$$

where γ_κ represents client importance weights.

4.3. Efficiency Analysis

Table 1 presents a comparative analysis of three federated aggregation methods: FedAvg, FedStack, and pFedMxF. In terms of computational characteristics, both pFedMxF and FedAvg maintain constant memory usage of $\mathcal{O}(1)$ when the number of clients increases. However, FedAvg lacks precise aggregation, requiring $\mathcal{O}(Krd)$ communication and $\mathcal{O}(rd)$ time complexity. FedStack achieves precise aggregation at the cost of increased resource demands, using $\mathcal{O}(K)$ memory, $\mathcal{O}(K^2rd)$ communication, and $\mathcal{O}(Krd)$ time complexity. Notably, pFedMxF emerges as an efficient alternative, combining $\mathcal{O}(1)$ memory efficiency with precise aggregation while achieving superior

communication complexity $\mathcal{O}(K|\Omega|)$ and time complexity $\mathcal{O}(|\Omega|\log|\Omega|)$. In pFedMxF, the number of trainable parameters $|\Omega|$ is significantly smaller than the trainable parameters of FedAvg. Given $d = 768$, pFedMxF with $|\Omega| = 3000$ has almost identical training parameters to FedAvg with $r = 2$, while achieving performance comparable to FedAvg with $r = 16$.

5. Experimental Evaluation

Datasets. We use three datasets: CIFAR-100 [23], TinyImageNet [24] and DomainNet [34] (in Appendix A). For a fair comparison with baseline class-incremental learning methods [1, 9, 17, 36, 38, 49] in the FCIL setting, we follow the same protocols proposed by [36] to set incremental tasks and utilize the identical class order generated from [36] and [28]. The local dataset for each client is generated under two types of non-i.i.d settings [25]: quantity-based label imbalance and distribution-based label imbalance. The degree of heterogeneity for these two settings is controlled by hyperparameters α and β. We run our experiments three times with different random seeds, and report both the final task performance and the averaged accuracy of all tasks.

Baselines. We compare our method with existing FCIL methods, including TARGET [58], GLFC [6] and LGA [7]. Additionally, we adapt several CIL methods, including EWC [22], LwF [27], iCaRL [36], L2P [46] for the FL setting. We also compare with two recently proposed LoRA-based CIL methods with orthogonal constraint (i.e., InfLoRA [28] and PILoRA [13]), implementing them in the FL setting. We also establish performance bounds using non-CL approaches. *Joint* serves as the naive approach (upper bound) that trains a base model without incrementation of classes. We evaluate the performance of these methods under various non-IID settings. For a fair comparison, all methods are fine-tuned from the same pre-trained model as ours.

Implementation. We adopt the self-supervised pretrained backbone (DINO [4]) for ViT-B/16 [8], which is widely used in CIL. Following the typical setup, the adapter is only inserted in the query and key linear layer in the attention block of the transformer. We set the number of frequency as $|\Omega| = 3000$ for pFedMxF and the LoRA rank $r = 16$ for FedAvg and FedStacking, InfLoRA and PILoRA. We train our models using Adam [20] with a batch size of 64 and following [57], we use different learning rates: $1e^{-3}$ for the classification layer and $1e^{-5}$ for the adapter parameters. Moreover, we also use cosine annealing in the training process. We set $\delta = 1, \lambda = 0.001, \gamma = 0.5$ and $\eta = 0.2$. We initialize 10 local clients to train and upload the parameters at each communication round. Each FL training epoch consists of 5 communication rounds, and in each global round, we randomly select 10 clients to conduct the local training.

Table 2. Test Accuracy (%) on CIFAR-100. Results are for 10 tasks (10 classes / task) under 2 non-IID settings.

(a) Quantity-Based Label Imbalance (QBLI)

Non-IID	QBLI					
Partition	$\alpha = 6$		$\alpha = 4$		$\alpha = 2$	
Methods	**Last**	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**
Joint	88.6	-	84.3	-	79.8	-
EWC	57.9	69.1	55.9	66.8	42.2	52.7
LwF	57.4	68.8	55.1	66.7	40.8	52.9
iCaRL	35.8	56.5	37.1	58.9	43.4	55.3
L2P	63.4	65.1	59.0	58.2	2.6	5.6
TARGET	60.9	71.3	58.8	69.5	45.2	56.5
GLFC	58.2	70.4	53.7	65.9	13.1	37.7
LGA	64.5	73.6	61.1	70.5	21.6	40.9
PILoRA	69.3	78.5	65.3	74.4	54.6	62.8
InfLoRA	70.5	78.4	66.7	75.6	56.3	62.5
pFedMxF	**71.3**	**80.7**	**67.4**	**76.2**	**57.0**	**64.9**

(b) Distribution-Based Label Imbalance (DBLI)

Non-IID	DBLI					
Partition	$\beta = 0.5$		$\beta = 0.1$		$\beta = 0.05$	
Methods	**Last**	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**
Joint	90.1	-	87.8	-	85.9	-
EWC	65.5	77.8	57.8	73.2	43.5	59.2
LwF	64.7	77.5	54.6	63.3	45.7	64.5
iCaRL	51.3	67.7	50.1	65.9	44.6	63.0
L2P	53.9	51.6	62.9	71.4	38.7	32.2
TARGET	66.1	77.8	60.5	71.1	51.8	65.3
GLFC	68.2	75.7	55.4	67.9	20.1	47.9
LGA	70.5	78.5	63.3	72.5	27.6	50.8
InfLoRA	68.4	78.4	63.3	73.8	54.2	67.5
PILoRA	**70.5**	78.2	63.0	73.5	57.5	69.3
pFedMxF	70.2	**80.3**	**65.6**	**75.2**	**60.5**	**70.5**

5.1. Main Results and Discussion

Table 2 and 3 demonstrate that pFedMxF effectively addresses the heterogeneity challenges in federated class-incremental learning across different scales of datasets (CIFAR-100 and TinyImageNet) under various Non-IID settings. Under quantity-based label imbalance (QBLI), pFedMxF consistently outperforms baseline methods, achieving strong accuracy on both datasets and maintaining robust performance even under severe imbalance ($\alpha = 2$), where methods like L2P fail catastrophically (dropping to $2.6/5.6\%$ on CIFAR-100 and $8.2\%/10.2\%$ on TinyImageNet). This supports our theoretical insight that frequency domain decomposition effectively addresses spatial heterogeneity by assigning orthogonal frequency components to different clients and tasks.

Resilience to spatial-temporal heterogeneity figure 3 and 4 comprehensively demonstrate pFedMxF's superior resilience to both spatial and temporal heterogeneity. For temporal heterogeneity (figure 3), we track accuracy across sequential tasks ($t = 0$ to 10) on both CIFAR-100 and TinyImageNet datasets. pFedMxF maintains consistently higher accuracy throughout the task sequence, showing sig-

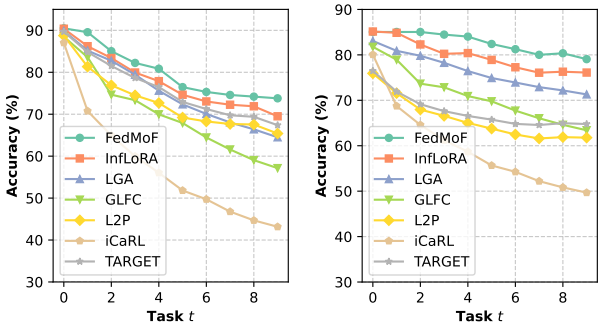
Table 3. Test Accuracy (%) on TinyImageNet. Results are for 10 tasks (10 classes / task) under 2 non-IID settings.

(a) Quantity-Based Label Imbalance

Non-IID	QBLI					
Partition	$\alpha = 6$		$\alpha = 4$		$\alpha = 2$	
Methods	**Last**	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**
Joint	83.6	-	82.9	-	80.2	-
iCaRL	51.3	72.4	51.8	60.3	45.8	56.9
L2P	61.6	58.0	49.4	39.3	8.2	10.2
TARGET	72.6	81.6	70.3	79.6	63.8	73.5
GLFC	69.1	77.9	61.3	73.5	25.1	39.4
LGA	71.3	79.4	65.8	75.3	36.7	48.8
InfLoRA	75.5	81.7	74.4	81.4	67.4	75.3
PILoRA	74.8	81.5	**74.7**	80.7	70.7	77.6
pFedMxF	**76.3**	**82.7**	74.4	**83.1**	**71.8**	**78.4**

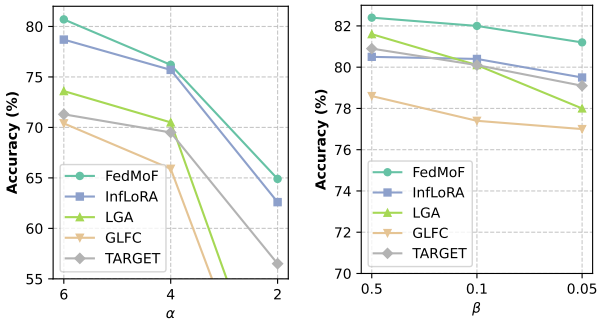
(b) Distribution-Based Label Imbalance

Non-IID	DBLI					
Partition	$\beta = 0.5$		$\beta = 0.1$		$\beta = 0.05$	
Methods	**Last**	**Avg.**	**Last**	**Avg.**	**Last**	**Avg.**
Joint	84.3	-	83.3	-	82.8	-
iCaRL	56.4	77.4	60.4	71.0	46.7	57.8
L2P	64.2	66.9	56.3	52.5	43.2	51.9
TARGET	71.6	80.9	71.0	80.1	69.3	79.1
GLFC	70.7	78.6	69.8	77.4	50.2	77.0
LGA	73.7	81.6	70.8	80.1	68.4	78.0
InfLoRA	74.3	80.6	74.3	81.1	72.9	79.8
PILoRA	74.6	81.3	74.2	79.9	73.1	80.3
pFedMxF	**76.2**	**82.4**	**76.1**	**82.3**	**74.5**	**81.9**



(a) CIFAR-100 (b) TinyImageNet

Figure 3. Investigation of anti-**temporal heterogeneity** in terms of Accuracy score on CIFAR-100 and TinyImageNet on QBLI setting



(a) CIFAR-100 (b) TinyImageNet

Figure 4. Investigation of anti-**spatial heterogeneity** in terms of Accuracy score on CIFAR-100 on QBLI and DBLI Non-IID setting.

nificantly less performance degradation compared to both typical FCIL methods (TARGET, GLFC, LGA) and LoRA-based approaches (InfLoRA). This superior temporal stability validates our theoretical insight that assigning orthogonal frequency components to different tasks effectively prevents catastrophic forgetting and interference between sequential updates. For spatial heterogeneity (figure 4), we evaluate performance under increasingly severe data distribution skews (QBLI: α from 6 to 2; DBLI: β from 0.5 to 0.05). pFedMxF demonstrates remarkable robustness, maintaining higher accuracy compared to baselines as heterogeneity increases. Notably, while competing methods show sharp performance drops under severe heterogeneity (particularly at $\alpha = 2$ and $\beta = 0.05$), pFedMxF's performance degrades more gracefully. This resilience to spatial heterogeneity validates our frequency-based decomposition strategy's effectiveness in handling non-IID data distributions.

5.2. Ablation Analysis

Comparison of aggregation methods. Table 6 demonstrates how pFedMxF effectively addresses the fundamental limitations of existing LoRA-based aggregation methods. While FedAvg reduces resource requirements through parameter averaging, it suffers from inaccurate aggregation between client updates equation 6, leading to degraded performance (77.8%, 73.2%, 61.5% for $\alpha = 6, 4, 2$ in QBLI). FedStack achieves exact aggregation through parameter stacking (equation 7) but at the cost of increased memory and communication overhead that scales with client numbers K, showing only modest improvements (78.6%, 74.4%, 62.6%). In contrast, pFedMxF successfully overcomes both limitations through frequency domain decomposition, achieving superior performance across all heterogeneity settings while maintaining constant memory usage and communication efficiency. This performance advantage is particularly evident under severe heterogeneity, validating that frequency-domain aggregation effectively addresses both the interference and scaling limitations of existing LoRA-based approaches while preserving their efficiency benefits.

Effectiveness of orthogonality. Table 5 provides compelling empirical evidence for the importance of orthogonal frequency assignments in pFedMxF's design. When frequency coordinates are shared across clients ("pFedMxF + Shared"), causing subspace overlap and disabling orthogonality, the performance significantly degrades compared to our proposed random non-overlapping frequency assignment ("pFedMxF + random") across all heterogeneity settings. This performance gap widens notably under severe hetero-

Table 4. A comparison of different classifiers. We report the **Average Accuracy** for all tasks in CIFAR-100 .

Non-IID	**QBLI** (α)			**DBLI** (β)		
Degree	6	4	2	0.5	0.1	0.05
pFedMxF + **Uni**	74.1	70.4	58.2	74.0	69.2	63.8
pFedMxF + **AAC**	80.7	76.2	64.4	80.3	75.2	69.0

Table 5. Shared *vs.*, non-shared frequency coordinates . We report the **Average Accuracy** for all tasks in CIFAR-100.

Non-IID	**QBLI** (α)			**DBLI** (β)		
Degree	6	4	2	0.5	0.1	0.05
pFedMxF + **Shared**	78.2	74.3	62.2	77.9	73.6	67.2
pFedMxF + **Random**	80.7	76.2	64.4	80.3	75.2	69.0

geneity conditions (QBLI $\alpha = 2$), directly validating our theoretical insight that assigning orthogonal frequency components to different clients prevents interference between updates while preserving perfect reconstruction capability. The results conclusively demonstrate that the orthogonality achieved through our frequency assignment strategy is instrumental to pFedMxF's superior performance in addressing both spatial and temporal heterogeneity challenges.

Effectiveness of AAC. Table 4 demonstrates the clear advantages of our proposed Auto-task Agnostic Classifier (AAC) over a unified classifier approach in handling heterogeneous federated class-incremental learning. While a unified classifier ("pFedMxF + Uni") that processes all classes through a joint probability distribution achieves moderate performance, our AAC design ("pFedMxF + AAC") significantly improves accuracy across all heterogeneity settings. This substantial improvement validates our theoretical design of decomposing the classification problem into task-specific classification and automatic routing, effectively preventing task interference through separate classifiers $\varphi^t(\cdot)$

Frequency vs. sparsity. Table 7 investigates the relationship between frequency component sparsity and model performance, demonstrating pFedMxF's efficiency in parameter utilization. Under both QBLI and DBLI settings, we vary the number of frequency components $|\Omega| \in (1000, 3000, 6000)$, resulting in different sparsity ratios $|\Omega|/d^2$ (from 1.6‰to 9.6‰). For a fair comparison, we match these configurations with equivalent LoRA ranks $r = (1, 2, 4)$ in terms of parameter count. The results show that pFedMxF achieves strong performance even with extremely sparse frequency components (80.2% accuracy with just 1.6‰sparsity), and reaches optimal performance (80.7%) at moderate sparsity (3.2‰). Most notably, pFedMxF with $|\Omega| = 3000$ (3.2‰sparsity) achieves higher performance comparable to LoRA rank $r = 4$, despite using only half the parameter.

Also, as visualized in figure 5, sparse frequency patterns

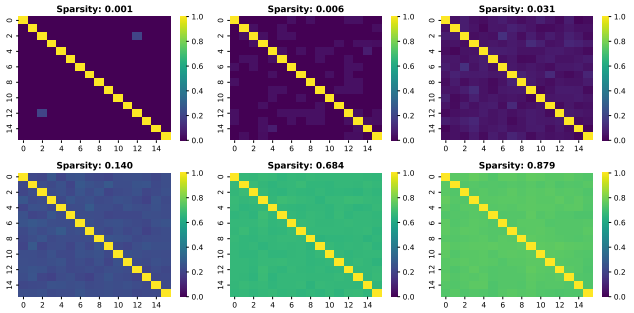


Figure 5. Visualization of cross-correlation between 16 different parameter updates $\Delta\mathbf{W}_\kappa^t$ at varying sparsity levels. Each matrix element represents the inner product $\langle \Delta\mathbf{W}_\kappa^t, \Delta\mathbf{W}_{\kappa'}^{t'} \rangle$ between pairs of updates. Lower values (darker colors) indicate stronger orthogonality between updates $\Delta\mathbf{W}_\kappa^t$, demonstrating how sparsity in frequency assignment maintains orthogonal subspaces.

Table 6. A comparison of aggregation methods . We report the **Average Accuracy** for all tasks in **CIFAR-100** under two non-IID settings.

Non-IID	**QBLI** (α)			**DBLI** (β)		
Partition	6	4	2	0.5	0.1	0.05
FedAvg	77.8	73.2	61.5	77.7	73.2	66.6
FedStack	78.6	74.4	62.6	78.1	73.7	67.1
pFedMxF	**80.7**	**76.2**	**64.4**	**80.3**	**75.2**	**69.0**

Table 7. Performance comparison under different frequency sparsity levels. pFedMxF achieves optimal performance with only 3.2‰ of the frequency space.

Non-IID	**QBLI** ($\alpha = 6$)			**DBLI** ($\beta = 0.5$)		
# Freq. $\lvert\Omega\rvert$	1000	3000	6000	1000	3000	6000
Sparsity $\lvert\Omega\rvert/d^2$	1.6‰	3.2‰	6.4‰	1.6‰	3.2‰	6.4‰
Equal rank r	1	2	4	1	2	4
pFedMxF	**80.2**	**80.7**	**80.6**	**79.1**	**80.3**	**80.5**
FedStack	76.8	77.4	77.25	76.7	77.3	78.5

clearly separate between clients and tasks (We randomly chose 16 $\Delta\boldsymbol{W}_\kappa^t$ from the training process). This low sparsity is sufficient to achieve strong performance while maintaining orthogonality between client updates.

6. Conclusions

We presented pFedMxF, a novel framework for FCIL that addresses spatial, temporal, and resource heterogeneity through orthogonal frequency component decomposition, enabling interference-free learning while maintaining a minimal communication overhead. pFedMxF achieves strong performance and orthogonality even with extremely sparse frequency components, establishing it as a practical approach that effectively balances performance, communication efficiency, and robustness to heterogeneity in FCIL.

Acknowledgements

The research is also supported, in part, by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL); the Ministry of Education, Singapore, under its Academic Research Fund Tier 1; and the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No. AISG2-RP-2020-019).

References

[1] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *ICCV*, pages 844–853, 2021.

[2] Gaurav Bagwe, Xiaoyong Yuan, Miao Pan, and Lan Zhang. Fed-cprompt: Contrastive prompt for rehearsal-free federated continual learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

[6] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10164–10173, 2022.

[7] Jiahua Dong, Yang Cong, Gan Sun, Yulun Zhang, Bernt Schiele, and Dengxin Dai. No one left behind: Real-world federated class-incremental learning. *arXiv preprint arXiv:2302.00903*, 2023.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pages 86–102, 2020.

[10] Tao Fan, Hanlin Gu, Xuemei Cao, Chee Seng Chan, Qian Chen, Yiqiang Chen, Yihui Feng, Yang Gu, Jiaxiang Geng, Bing Luo, et al. Ten challenging problems in federated foundation models. *arXiv preprint arXiv:2502.12176*, 2025.

[11] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. *arXiv preprint arXiv:2303.10070*, 2023.

[12] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3614–3631, 2020.

[13] Haiyang Guo, Fei Zhu, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. Pilora: Prototype guided incremental lora for federated class-incremental learning. In *Proceedings of the European Conference on Computer Vision*, 2024.

[14] 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.

[15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[17] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *CVPR*, 2021.

[18] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87, 1991.

[19] Latif U Khan, Walid Saad, Zhu Han, Ekram Hossain, and Choong Seon Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, 23(3):1759–1799, 2021.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526, 2017.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[24] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[25] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.

[26] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future

directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[28] Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.

[29] Chenghao Liu, Xiaoyang Qu, Jianzong Wang, and Jing Xiao. Fedet: A communication-efficient federated class-incremental learning framework based on enhanced transformer. *arXiv preprint arXiv:2306.15347*, 2023.

[30] 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.

[31] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989.

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[33] Solmaz Niknam, Harpreet S Dhillon, and Jeffrey H Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6):46–51, 2020.

[34] 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.

[35] 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.

[36] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[37] Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5054–5062, 2023.

[38] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. On learning the geodesic path for incremental learning. In *CVPR*, 2021.

[39] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

[40] Zixing Song, Yifei Zhang, and Irwin King. No change, no gain: Empowering graph neural networks with expected model change maximization for active learning. In *NeurIPS*, 2023.

[41] Zixing Song, Yifei Zhang, and Irwin King. Optimal block-wise asymmetric graph construction for graph-based semi-supervised learning. In *NeurIPS*, 2023.

[42] Zixing Song, Ziqiao Meng, and Irwin King. A diffusion-based pre-training framework for crystal property prediction. In *AAAI*, pages 8993–9001. AAAI Press, 2024.

[43] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[44] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*, 2023.

[45] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022.

[46] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[47] Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*, 2024.

[48] Xiwen Wei, Guihong Li, and Radu Marculescu. Online-LoRA: Task-free Online Continual Learning via Low Rank Adaptation. In *NeurIPS 2024 Workshop on Scalable Continual Learning for Lifelong Foundation Models*, 2024.

[49] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *CVPR*, pages 374–382, 2019.

[50] Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, and Rex Ying. Low-Rank Adaptation for Foundation Models: A Comprehensive Review, 2024.

[51] Qiang Yang, Lixin Fan, and Han Yu. *Federated Learning: Privacy and Incentive*. Springer, Cham, 2020.

[52] Shu Yang, Muhammad Asif Ali, Cheng-Long Wang, Lijie Hu, and Di Wang. MoRAL: MoE Augmented LoRA for LLMs' Lifelong Learning, 2024.

[53] Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pFedLoRA: Model-heterogeneous personalized federated learning with LoRA tuning. *arXiv preprint arXiv:2310.13283*, 2023.

[54] Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S. Yu, and Irwin King. Recent Advances of Multimodal Continual Learning: A Comprehensive Survey, 2024.

[55] Han Yu, Xiaoxiao Li, Zenglin Xu, Randy Goebel, and Irwin King. *Federated Learning in the Age of Foundation Models*. Springer Cham, 2025.

[56] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 23219–23230. IEEE, 2024.

[57] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. *arXiv preprint arXiv:2303.05118*, 2023.

[58] Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lv. Addressing catastrophic forgetting in federated class-continual learning. *arXiv preprint arXiv:2303.06937*, 2023.

[59] Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023.

[60] Yifei Zhang, Hao Zhu, Aiwei Liu, Han Yu, Piotr Koniusz, and Irwin King. Less is More: Extreme Gradient Boost Rank-1 Adaption for Efficient Finetuning of LLMs, 2024.

[61] 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.

[62] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.