Fisher-Routed Mixture of Experts for Federated Class-Incremental Learning

Wenhao Yuan[1], Chenchen Lin[2], Jian Chen[1], Jinfeng Xu[1], Zewei Liu[1], and Edith Cheuk Han Ngai[1,*]

[1] Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong SAR, China
[2] School of Artificial Intelligence, Sun Yat-sen University, China.
wenhao.yuan@connect.hku.hk, chngai@eee.hku.hk

Abstract. Federated Learning (FL) emerged as a promising distributed machine learning paradigm. However, extending FL to the class incremental learning scenarios introduces unique challenges: 1) Capacity conflict and catastrophic forgetting from the shared model overloading, 2) Heterogeneity from Non-Independent and Identically Distributed (Non-IID) data, and 3) Synchronized class misalignment. In this paper, we propose **F**isher-Routed **MiX**ture of Experts for **Fed**erated Class-Incremental Learning (FedFMX), a novel framework to address these challenges via adaptive expert specialization across clients. The crucial insight is to route each sample to an expert subset that jointly optimizes knowledge acquisition and retention. Specifically, we introduce a Fisher-Routed Expert Scoring (FRES) module to estimate expert importance via Fisher-based stability cost and gradient-based plasticity gain. Then, we design an Adaptive Expert Selection (AES) module by quantifying marginal contributions for adaptive expert subset determination. Finally, by the routing-aware regularization (RAR), we achieve load balance and efficient FL training. We theoretically prove the $\mathcal{O}(T^{-1})$ convergence rate. Extensive experiments on multiple benchmarks compared with state-of-the-art methods demonstrate the superiority of FedFMX.

Keywords: Federated Class-Incremental Learning · Fisher Information Matrix · Mixture of Experts

1 Introduction

Federated Learning (FL) [8, 21, 29, 42] has emerged as a promising distributed learning paradigm, enabling multiple geographically decentralized clients to jointly learn a shared global model without exposing their raw data. However, conventional FL approaches typically assume that *the label distribution remains static and consistent during the training process* [38, 71], which rarely holds in real-world applications [10, 14, 62]. In practice, due to dynamic contexts, client data

* Corresponding author.

distributions evolve, and clients continuously receive new classes of data with previously unseen classes [6,53], thus invalidating the aforementioned ideal assumption. Further, given that clients are generally resource-constrained with limited storage and computational capacity [34], maintaining a complete history of past data is infeasible, which exacerbates the risk of catastrophic forgetting [1,67].

To overcome the limitation of the static label distribution assumption, recent efforts have advanced *Federated Class-Incremental Learning* (FCIL) paradigm [5, 10], enabling clients to incrementally learn novel classes while collaboratively updating the global model [38]. However, the intersection of the federated and incremental paradigms introduces several intrinsic and crucial challenges [6,71]: (i) *Capacity Conflict and Catastrophic Forgetting*: Since the shared global model is incrementally overloaded with heterogeneous data, the learned representations are overwritten, leading to catastrophic forgetting and degraded generalization [23,65]; (ii) *Statistical Heterogeneity*: Clients possess dynamically evolving Non-IID data, resulting in the gradient direction inconsistency and biased local updates, impairing global convergence [13,16]; (iii) *Synchronized Class Misalignment*: Clients encounter newly introduced classes at various phases, resulting in temporally inconsistent label spaces, exacerbating model drifting [15,71].

To alleviate catastrophic forgetting in FCIL, prior works adopt global regularization, memory replay, or rehearsal-free techniques [6,55,67]. However, these methods default the global model to a monolithic entity, without fine-grained mechanisms to modulate stability and plasticity, particularly pronounced in heterogeneous scenarios that require differentiated treatment across model parameters. Drawing from these analyses, we raise **Key Question I**: *How to balance stability and plasticity in FCIL such that the global model retains prior knowledge while adapting to incrementally emerging classes across heterogeneous clients?* Another fundamental challenge arises from the compounding effect of statistical heterogeneity and temporal class drifting. Clients dynamically receive novel classes at different phases, leading to the *synchronized class misalignment* problem, where the semantic inconsistency of class distributions across clients impairs convergence stability and generalization [6,67,71]. While recent works introduce personalized methods to mitigate heterogeneity [64], overlooking tailoring collaboration based on clients' evolving task profiles, limits the scalability. Then, we pose **Key Question II:** *How to achieve stable and effective collaboration in FCIL under client heterogeneity and temporally divergent class updates, ensuring robust and efficient training?*

To address these issues, we present a novel solution, **F**isher-Routed **MiX**ture of Experts for **Fed**erated Class-Incremental Learning (FEDFMX), which introduces the expert modularity and adaptive activation and routing to disentangle evolving and heterogeneous client dynamics. For question **I**, we design a *Fisher-Routed Expert Scoring* (FRES) module to estimate suitability by quantifying the stability cost and plasticity gain of each expert through Fisher information in § 3.2, enabling context-aware routing. For question **II**, we introduce an *Adaptive Expert Selection* (AES) module in § 3.3, which formulates expert subset selection as a cooperative game and activates experts dynamically via the marginal contri-

bution to stability–plasticity trade-offs. To enhance routing robustness and mitigate biased expert utilization, we present a *routing-aware regularization* (RAR) scheme in § 3.4. Our main contributions are summarized as follows:

- We identify the significant challenges in FCIL and propose FedFMX to mitigate catastrophic forgetting and heterogeneity by routing each sample to a dynamical expert subset, achieving fine-grained specialization and robust knowledge integration.
- We design the FRES module to quantify the expert suitability via stability-plasticity trade-offs, where we evaluate the marginal contribution of each expert and determine adaptive expert subsets, enabling efficient expert collaboration across temporally misaligned classes. Through the RAR strategy, we promote stable and fair expert utilization.
- Extensive experiments on CIFAR-10, CIFAR-100 [25], and Tiny-ImageNet [26] datasets validate the effectiveness of FedFMX, showing consistent improvements over state-of-the-art methods.

2 Related Work

2.1 Federated Class-Incremental Learning

Federated Class-Incremental Learning (FCIL) [6] extends the FL paradigm to continual learning with training while incrementally acquiring new classes [13, 36], which poses crucial challenges, such as asynchronous task arrivals [23,31,71] and Non-IID data [38,59], leading to catastrophic forgetting and unbalanced updates [10,65,67]. To solve these issues, regularization-based methods [10,56] are proposed to constrain updates to preserve previously acquired knowledge. For instance, CGoFed [10] and LGA [5] include group-aware regularization to alleviate catastrophic forgetting. Meanwhile, many approaches adopt replay-based strategies [23, 33, 46, 55] to maintain a local memory buffer or a synthetic generator to replay past samples. In contrast, rehearsal-free methods [17, 35, 55] eliminate memory storage by relying on parameter isolation or prompt tuning to preserve prior knowledge. Personalization [40, 60] and adaptive [66, 72] methods address heterogeneity through task-dependent model adaptation [27] or dynamic aggregation [47]. pFedMxF [71] and SpaPFCIL [72] balance personalization and task continuity by selectively combining local adaptation with global coordination. Nevertheless, most existing FCIL methods rely on conventional naive parameter aggregation strategies, tending to overwrite the knowledge boundaries of previous classes, leading to unstable performance across phases and clients.

2.2 MoE-based Federated Learning

Mixture-of-Experts (MoE) [41, 68, 73] has emerged as a powerful architecture to address the challenges of client heterogeneity and scalability in FL [11,61,63,74], which enhances model capacity through expert modularization and task routing [39, 43]. Nonetheless, integrating MoE into FL introduces unique challenges,

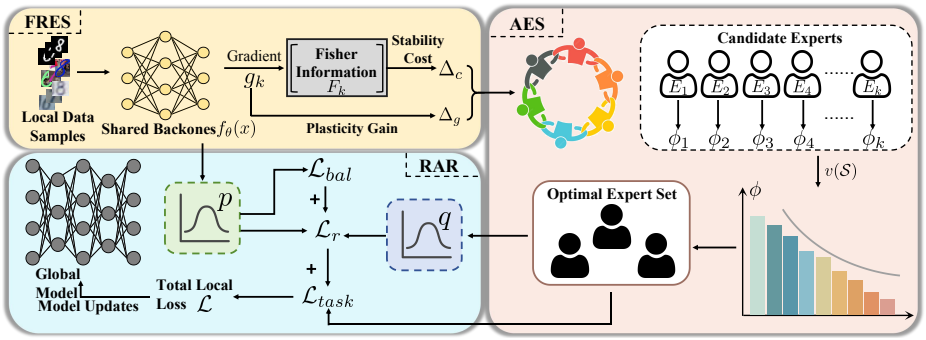

Fig. 1: Overview illustration of the proposed FedFMX framework.

such as failing to adapt to heterogeneous clients and evolving tasks [3], imbalanced loading for dominant experts [4,50]. To address these issues, recent works have explored MoE-based aggregation [43,69], which employs globally shared expert pools with the activation per client or task. Dynamic routing and enhanced gating methods [9,44,51] select experts based on gradient or reward feedback. Personalized MoE structures [11,63,74] provide client-specific sub-expert sets to enhance specialization. Despite these advancements, existing MoE-based methods fail to reconcile adaptability with knowledge retention in class-incremental settings and accommodate asynchronous task arrivals, motivating the need for a framework that routes knowledge dynamically without compromising previously acquired representations.

3 Methodology

In this section, we present the detailed methodology of FedFMX. We begin with the preliminaries of FCIL in § 3.1 and elaborate on the FRES module design in § 3.2. We describe the AES module with the marginal contribution metric in § 3.3. Finally, we present the route-aware regularization in § 3.4. The overall framework is shown in Figure 1

3.1 Preliminaries

We extend the standard *class-incremental learning* (CIL) to federated settings. Given N distributed clients that collaboratively and incrementally train a shared global model, each client $i \in \mathcal{I} = \{1, \ldots, N\}$ receives a sequence of streaming tasks $\{\mathcal{T}_i^{(t)}\}_{t=1}^{T}$, where $\mathcal{T}_i^{(t)} = \{\mathcal{X}_i^{(t)}, \mathcal{Y}_i^{(t)}\}$ induces a local dataset $\mathcal{D}_i^{(t)} = \{(\boldsymbol{x}_{i,j}^{(t)}, y_{i,j}^{(t)})\}_{j=1}^{n_i^{(t)}}$, where $n_i^{(t)} = |\mathcal{D}_i^{(t)}|$, the input samples $\boldsymbol{x}_{i,j}^{(t)} \in \mathcal{X}$ and the labels $y_{i,j}^{(t)} \in \mathcal{Y}_i^{(t)} \subseteq \mathcal{Y}$ such that the labels in different tasks are disjoint per client across stages, i.e., $\mathcal{Y}_i^{(t)} \cap \mathcal{Y}_i^{(t')} = \varnothing$ for $t \neq t'$. For each client i, when training on $\mathcal{D}_i^{(t)}$,

the data from previous tasks $\{\mathcal{T}_i^{(k)}\}_{k=1}^{t-1}$ are inaccessible. Let $f(\cdot;\boldsymbol{w})$ denote the global model. The optimization objective can be characterized as:

$$\min_{\boldsymbol{w}} \frac{1}{D} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{n_i^{(t)}} \mathcal{L}(f(\boldsymbol{x}_{i,j}^{(t)};\boldsymbol{w}), y_{i,j}^{(t)}), \tag{1}$$

where $D = \sum_{t=1}^{T} \sum_{i=1}^{N} n_i^{(t)}$, and $\mathcal{L}$ denotes the task loss.

3.2 Fisher-Routed Expert Scoring Design

In FCIL, catastrophic forgetting and gradient interference are exacerbated by Non-IID data and asynchronous class arrivals across clients. Gradients from various tasks are entangled within a shared parameter space, limiting representational capacity. To mitigate such conflicts, we decompose the model into a shared backbone and multiple expert-specific parameter subspaces, enabling routing-conditioned optimization. Note that the expert subspaces are not statically tied to particular clients or classes; instead, they are dynamically activated at the sample level, facilitating adaptive and collaborative knowledge allocation across tasks. However, routing based solely on loss or fixed Top-K gating may cause over-specialization and amplify forgetting. To address this issue, we propose a *Fisher-Routed Expert Scoring* (**FRES**) module that formulates routing as a stability–plasticity trade-off. For each input sample, FRES constructs a Fisher-informed suitability score that jointly evaluates *plasticity gain*, reflecting an expert's capacity to absorb new information, and *stability cost*, quantifying the risk of disrupting curvature-sensitive directions associated with prior tasks.

Concretely, we instantiate the decomposition by partitioning the parameters into a shared backbone f_θ and K expert-specific subspaces $\{\Theta_k\}_{k=1}^{K}$, where each Θ_k comprises a dedicated body E_{φ_k} and classification head W_{ψ_k}, forming independent adaptation channels rather than merely attaching personalized heads. Given any input sample x, expert k produces logits $h_k(x) = W_{\psi_k}(E_{\varphi_k}(f_\theta(x)))$ with loss $\ell_k(x,y) = \mathrm{CE}(h_k(x), y)$ [45,52]. We approximate parameter importance via the *empirical Fisher information matrix* $\mathbb{E}[g_k g_k^\top]$, $g_k = \nabla_{\psi_k} \ell_k(x,y)$, which captures the local curvature of each expert subspace. Unlike classical consolidation methods that apply Fisher statistics as an update regularizer [24,57], we use them as a pre-update routing criterion to assess compatibility between incoming gradients and expert subspace. For efficiency, we adopt *diagonal approximation* and maintain an *exponential moving average* of squared gradients as:

$$F_k \leftarrow \rho F_k + (1-\rho)(g_k \odot g_k), g_k \triangleq \nabla_{\psi_k} \ell_k(x,y), \tag{2}$$

where $\rho \in [0,1)$ represents the decay rate and F_k captures the influence of the parameters of expert k in past predictions. Notably, we maintain diagonal Fisher statistics only for the classification heads and reuse gradients computed for backpropagation, without additional backward passes or second-order computations.

To ensure comparability, we apply the gradient normalization $\tilde{g}_k = g_k / \|g_k\|_2$ and standardize F_k per dimension and then apply a mapping to obtain $\widetilde{F}_k$.

Subsequently, we define two metrics to characterize expert k's interaction with the sample: (i) *Stability Cost.* The expected loss of prior knowledge is estimated:

$$\Delta_c^{(k)} = \frac{1}{2}\|\mathrm{diag}(\widetilde{F}_k)^{1/2}\tilde{g}_k\|_2^2, \tag{3}$$

which penalizes gradient directions salient in the current update yet important historically, based on Fisher statistics accumulated before the current sample. (ii) *Plasticity Gain.* The potential improvement from learning the sample is:

$$\Delta_g^{(k)} = \|g_k\|_2^2/(\mathrm{Tr}(\widetilde{F}_k) + \varepsilon), \tag{4}$$

where the constant ε is for numerical stability. $\Delta_g^{(k)}$ approximates the effective learning gain by balancing the gradient magnitude against local parameter rigidity. Here, $\|g_k\|_2^2$ represents the first-order learning signal induced by the sample, while $\mathrm{Tr}(\widetilde{F}_k)$ captures the curvature and historical importance of the expert parameters. Thus, $\Delta_g^{(k)}$ favors experts that can achieve substantial loss reduction while operating in relatively flat parameter regions. Larger $\Delta_g^{(k)}$ values therefore indicate stronger and safer responsiveness to new information.

Unlike conventional dynamic head or expert selection strategies that primarily rely on instantaneous loss or gating scores, FRES calculates $\Delta_c^{(k)}$ and $\Delta_g^{(k)}$ before activation for each candidate expert during the training phase to evaluate trade-offs between learning potential and forgetting risk, enabling the routing layer to avoid over-specialized or fragile experts.

3.3 Adaptive Experts Selection

To enable context-aware and sample-specific expert activation, we propose an *Adaptive Expert Selection* (**AES**) module. In contrast to the conventional static Top-1 or Top-K gating method, which fails to adapt to evolving task demands or learning phases, AES dynamically determines the optimal subset of experts $\mathcal{K}_b$ for each input sample x_b from the batch $\{(x_b, y_b)\}_{b=1}^B$. The selection process explicitly considers the *collaborative value* and *individual contribution*, promoting balanced specialization and avoiding the overfitting to dominant experts.

We formulate the expert selection as a cooperative game over the candidate expert set $\mathcal{C} = \{1, \ldots, K\}$. For any subset $\mathcal{S} \subseteq \mathcal{C}$, we define the value function $v(\mathcal{S})$ to quantify the utility of activating experts in $\mathcal{S}$ for any sample (x, y):

$$v(\mathcal{S}) = -\ell_{\mathcal{S}}(x, y) - \alpha \sum\nolimits_{k\in\mathcal{S}} \Delta_c^{(k)} + (1-\alpha)\overline{\Delta}_g(\mathcal{S}), \tag{5}$$

where $\alpha \in [0, 1]$ governs the balance between prediction accuracy and forgetting risk; $\ell_{\mathcal{S}}(x, y) = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \ell_k(x, y)$ denotes the averaged loss under expert subset $\mathcal{S}$; $\overline{\Delta}_g(\mathcal{S}) = \frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \tilde{\Delta}_g^{(k)}$ with $\tilde{\Delta}_g^{(k)}$ as the normalized $\Delta_g^{(k)}$. (5) captures the dual goals of minimizing loss and maximizing gain, while accounting for the potential risk of disrupting previously acquired knowledge. To allow incremental $\mathcal{O}(1)$ updates when expanding $\mathcal{S}_{j-1}$ to $\mathcal{S}_j$, we maintain cached aggregates

$\{\sum_{\ell}, \sum_{\Delta_c}, \sum_{\tilde{\Delta}_g}, |\mathcal{S}|\}$ with $\sum_{\Delta_c}(\mathcal{S}) = \sum_{k\in\mathcal{S}} \Delta_c^{(k)}$, $\sum_{\tilde{\Delta}_g}(\mathcal{S}) = \sum_{k\in\mathcal{S}} \tilde{\Delta}_g^{(k)}$. The individual marginal contribution of each expert $k \in \mathcal{C}$ is given by:

$$\phi_k = \sum\nolimits_{\mathcal{S}\subseteq\mathcal{C}\setminus\{k\}} \frac{|\mathcal{S}|!(K-|\mathcal{S}|-1)!}{K!}[v(\mathcal{S}\cup\{k\}) - v(\mathcal{S})], \tag{6}$$

which quantifies the expected marginal contribution of each expert k over all potent coalitions. Then, to choose the activated expert set $\mathcal{K}_b$ for sample x_b, we sort all experts in descending order of their marginal contribution values by $\phi_{k_{(1)}} \geq \cdots \geq \phi_{k_{(K)}}$. Define $\mathcal{S}_0 = \varnothing$, we iteratively construct subsets $\mathcal{S}_j = \mathcal{S}_{j-1} \cup \{k_{(j)}\}$ for $j \in \mathcal{C}$, and compute the marginal coalition gain:

$$\Delta v_j = v(\mathcal{S}_j) - v(\mathcal{S}_{j-1}), \tag{7}$$

which terminates at the smallest index $j^\star \geq 2$ satisfying $\Delta v_j \leq \mu \overline{\Delta v}_{1:j-1}$ with $\overline{\Delta v}_{1:j-1} = \frac{1}{j-1}\sum_{p=1}^{j-1} \Delta v_p$, where $\mu \in [0,1]$ controls the strictness of diminishing gain. A smaller μ leads to earlier stopping and fewer experts. The final activation set for the sample x_b is given by $\mathcal{K}_b = \mathcal{S}_{j^\star}$.

AES module formulates expert collaboration as a cooperative game, jointly accounting for individual contribution and coalition value to avoid greedy maximization of $v(\cdot)$, thereby mitigating expert over-concentration and forgetting. Since the expert pool in FCIL generally satisfies $K \leq 8$ [64], we compute exact Shapley values; for larger K, permutation-based estimators can be used instead.

3.4 Routing-aware Regularization

To guarantee stable and fair expert utilization and address following issues: (i) *Expert Over-concentration:* Routing decisions concentrated on a few dominant experts, leading to resource underutilization and training instability; (ii) *Stage-End Forgetting Risk:* During later incremental stages, the risk of forgetting learned classes remains high, we adopt the routing-aware regularization (**RAR**) strategy with a load-balancing regularization term and distilled gating mechanism, which extend the stability–plasticity trade-off from expert selection to optimization.

We first introduce a load-balancing regularization term to encourage smoother expert activation distributions and prevent overloading. Specifically, for a training batch of size B, we define the entropy-based regularizer as:

$$\mathcal{L}_{bal} = \frac{1}{B}\sum\nolimits_{b=1}^{B}\sum\nolimits_{k=1}^{K} p(k \mid x_b)\log p(k \mid x_b), \tag{8}$$

where $p(k \mid x_b)$ indicates the soft expert routing distribution for sample x_b. Minimizing the term $\mathcal{L}_{bal}$ discourages sharp routing distributions and mitigates the overload of minority experts. Then, to decouple routing from marginal contribution calculation while retaining the training-time benefit, we adopt a distillation-based regularization scheme. Specifically, we convert the expert importance scores ϕ_k in (6) into a soft *teacher* distribution $q(k \mid x_b) = \mathrm{softmax}(\phi_k/\tau_r)$ with temperature hyperparameter $\tau_r > 0$ controlling the sharpness. Subsequently, we instantiate a lightweight gating network g_η that maps the backbone

feature to logits $z^{(g)} \in \mathbb{R}^K$, which are converted into the *student* distribution as $p(k \mid x_b) = \text{softmax}(z_k^{(g)}/T)$ with distillation temperature $T > 0$. To ensure that the student network captures the expert activation patterns implied by the marginal contribution scores, the distillation loss is defined as:

$$\mathcal{L}_r = \frac{1}{B} \sum\nolimits_{b=1}^{B} \text{KL}(q(\cdot \mid x_b) \| p(\cdot \mid x_b)), \tag{9}$$

which regularizes the gating network to align with the semantics of Shapley-based expert selection while allowing efficient routing at inference. This distillation objective serves two key purposes: (i) it eliminates the need to compute costly marginal contributions during deployment, and (ii) it encourages the learned routing mechanism to generalize expert specialization patterns beyond the training distribution. By combining the above components into a unified loss function for local training. For each batch $\{(x_b, y_b)\}_{b=1}^{B}$, the total local loss is:

$$\mathcal{L} = \mathcal{L}_{task} + \beta \mathcal{L}_{bal} + \gamma \mathcal{L}_r, \tag{10}$$

where $\mathcal{L}_{task} = \frac{1}{B} \sum_{b=1}^{B} \sum_{k \in \mathcal{K}_b} w_k \ell_k$; β and γ control the strength of load balancing and route distillation. For training stability, we define the weight $w_k = \frac{\max(\phi_k, 0)}{\sum_{j \in \mathcal{K}_b} \max(\phi_j, 0)}$ to balance expert updates within the active set, allowing experts with strong positive marginal values to contribute, while suppressing noisy or redundant experts. We provide the detailed algorithm in Appendix A.

3.5 Discussion

FedFMX enables stability-aware training and efficient deployment via adaptive expert routing and distillation-based gating. During training, marginal contribution scores are transformed into a teacher distribution $q(\cdot \mid x_b)$, based on which AES selects a hard expert subset $\mathcal{K}_b$ for forward and backward propagation. A lightweight gating network produces a student distribution $p(\cdot \mid x_b)$, trained to match $q(\cdot \mid x_b)$ via KL distillation. The marginal contribution scores are used exclusively for expert subset selection and contribution-based weighting, and are treated as non-differentiable routing signals that do not participate in gradient updates. During inference, Top-1/Top-K experts are selected according to the distilled distribution $p(\cdot \mid x)$ and aggregated with normalized gating weights, enabling forward-only inference without marginal contribution computation.

4 Theoretical Analysis

We analyze convergence under standard federated optimization assumptions, including Lipschitz smoothness, unbiased gradients, and bounded first- and second-moment conditions (see Appendix B). AES induces an iteration-dependent projection onto activated expert subspaces, resulting in a projected-gradient FedAvg update. The shared backbone is updated using all samples, while the projection

applies only to expert-specific parameters. Although FRES determines the projection and regularizers modify the local objective, the overall update remains projected-gradient optimization. Following [22,54], we establish convergence under bounded residual gradient energy outside the activated subspace. Proofs are provided in Appendix C-E.

Assumption 1 *(Bounded Activated-Subspace Gradient Residual) The AES module induces an activation-dependent parameter subspace $\mathcal{S}_i(\boldsymbol{w}) \subseteq \mathbb{R}^d$ with projection operator $P_{\mathcal{S}_i(\boldsymbol{w})}$. The effective update under AES corresponds to the projected gradient $\hat{g}_i(\boldsymbol{w}) = P_{\mathcal{S}_i(\boldsymbol{w})} g_i(\boldsymbol{w})$. Assume that the discarded gradient energy outside the activated subspace is uniformly bounded, i.e., there exists $\delta \geq 0$ such that*

$$\mathbb{E}[\|(I - P_{\mathcal{S}_i(\boldsymbol{w})}) g_i(\boldsymbol{w})\|^2] \leq \delta^2, \forall i, \boldsymbol{w}. \tag{11}$$

Based on the assumptions, we can derive the following Lemma and Theorem.

Lemma 1. ***Local Training.*** *Consider the local update on client $i \in \{1, \ldots, N\}$ by $\boldsymbol{w}^+ = \boldsymbol{w} - \eta \hat{g}_i(\boldsymbol{w})$, for $\eta \leq \frac{1}{2\beta}$, the expected convergent upper bound of the local objective function for any consecutive global iterations is given by*

$$\mathbb{E}[F_i(\boldsymbol{w}^+) \mid \boldsymbol{w}] \leq F_i(\boldsymbol{w}) - \frac{\eta}{4}\|\nabla F_i(\boldsymbol{w})\|^2 + 2\sigma^2\beta\eta^2 + (\beta\eta^2 + \frac{\eta}{2})\delta^2. \tag{12}$$

Lemma 2. ***Local Drift.*** *Consider E local steps. Define $\boldsymbol{w}_i^{t,0} = \boldsymbol{w}^t$, and $\boldsymbol{w}_i^{t,e+1} = \boldsymbol{w}_i^{t,e} - \eta \hat{g}_i(\boldsymbol{w}_i^{t,e})$, for $e = \{0, \ldots, E-1\}$ and $\eta \leq \frac{1}{4\beta E}$, then*

$$\mathbb{E}[\|\boldsymbol{w}_i^{t,E} - \boldsymbol{w}^t\|^2 \mid \boldsymbol{w}^t] \leq 16\eta^2 E^2 (\|\nabla F_i(\boldsymbol{w}^t)\|^2 + \sigma^2 + \delta^2). \tag{13}$$

Theorem 1. ***Global Convergence.*** *Let $F(\boldsymbol{w}) \triangleq \sum_{i=1}^N \lambda_i F_i(\boldsymbol{w})$ with $\sum_{i=1}^N \lambda_i = 1$. For $\eta \leq \min\{\frac{1}{2\beta}, \frac{1}{4\beta E}, \frac{\kappa}{8\beta E \kappa_G^2}, \frac{\kappa^2}{48 E M_V}\}$ and any $T \geq 1$, it satifies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\boldsymbol{w}^t)\|^2] \leq \frac{8(F(\boldsymbol{w}^0) - F(\boldsymbol{w}^\star))}{\kappa\eta E T} + 6(2M + 3(\sigma^2 + \delta^2))(\frac{2}{\kappa^2} + \frac{1}{\kappa}). \tag{14}$$

In Theorem 1, the first term decays as $\mathcal{O}(1/T)$, while the second term forms an error floor determined by the gradient noise and the activated-subspace residual δ. In the deterministic and full-subspace case, the bound reduces to $\mathcal{O}(1/T)$.

5 Numerical Experiments

5.1 Experiment Setups

Datasets & Model. We evaluate our method on four real-world image classification tasks: CIFAR-10/CIFAR-100 [25], Tiny-ImageNet [26], and DomainNet [48]. We follow the experimental settings in [49,71] and adopt two backbone architectures for training: ResNet-18 [18] and ViT-B/16 [7] (in Appendix G),

Table 1: Accuracy comparison of FEDFMX and other benchmark methods on ResNet-18 backbone under the fine-grained setting. The best accuracy is in **bold**.

(a) Under the QLI setting

Method	CIFAR-10			CIFAR-100			Tiny-ImageNet			DomainNet		
	$\alpha_Q=2$	$\alpha_Q=4$	$\alpha_Q=6$	$\alpha_Q=2$	$\alpha_Q=4$	$\alpha_Q=6$	$\alpha_Q=2$	$\alpha_Q=4$	$\alpha_Q=6$	$\alpha_Q=2$	$\alpha_Q=4$	$\alpha_Q=6$
FEDMUT	18.52±2.2	23.41±2.6	28.96±2.7	5.43±0.9	7.76±1.4	9.85±1.8	2.96±0.8	4.10±0.6	5.21±0.7	4.72±0.9	6.84±1.2	8.93±1.1
FEDCROSS	20.07±2.1	25.48±2.3	31.03±2.2	6.79±1.3	9.18±1.6	11.24±1.7	3.41±0.6	4.66±0.7	5.84±0.6	5.61±1.1	7.93±1.4	10.24±1.2
FEDLSA	16.44±1.9	21.36±2.6	26.62±2.4	4.84±0.8	6.86±1.2	8.91±1.5	2.83±0.6	3.81±0.9	4.86±0.7	4.18±0.8	6.11±1.3	8.21±1.0
FEDWMSAM	19.61±2.3	24.62±2.1	29.57±2.8	6.52±1.9	8.85±1.5	10.83±1.6	3.24±0.6	4.46±0.6	5.62±0.7	5.23±0.9	7.58±1.1	9.82±1.6
PFEDMOE	41.47±2.2	47.96±2.1	53.56±2.5	24.83±1.4	30.76±1.3	34.82±1.1	8.79±0.5	11.02±0.8	12.18±0.7	21.76±1.4	27.92±1.2	32.18±1.3
FEDEWC	17.11±2.5	21.14±2.4	30.52±2.6	9.14±1.4	12.32±1.0	13.46±1.5	4.43±0.7	5.07±0.8	6.18±0.9	6.98±1.1	9.11±1.3	11.37±1.5
FEDLWF	35.93±2.9	39.25±2.1	52.47±2.9	20.12±1.1	26.09±1.2	29.34±1.3	4.85±0.9	7.14±0.6	11.22±0.7	18.42±1.3	23.96±1.5	27.85±1.2
TARGET	33.29±2.4	36.67±2.1	49.75±2.3	15.29±1.1	22.27±0.7	26.28±1.2	6.16±0.8	6.48±0.5	8.45±1.1	13.52±1.0	19.83±1.8	24.26±1.4
LANDER	36.68±2.5	40.49±2.4	59.33±2.4	28.28±1.6	35.11±1.1	38.45±1.6	9.22±0.9	9.53±0.7	11.23±0.8	24.17±1.6	30.85±1.3	34.94±1.1
RE-FED	51.96±3.1	55.13±2.9	59.58±2.5	30.55±1.5	40.28±0.9	41.52±1.5	9.97±0.6	12.06±0.6	13.37±0.7	26.94±1.7	36.18±1.3	38.42±1.6
FL-CLIP	50.25±2.9	53.26±2.7	59.47±2.4	29.76±1.6	39.15±1.2	40.87±1.4	9.51±0.8	11.85±1.0	13.15±0.6	25.83±1.1	34.77±1.6	37.15±1.3
FEDCBDR	57.62±2.6	61.45±2.0	65.79±1.9	45.54±1.4	46.97±1.2	48.15±1.3	13.44±0.7	15.32±0.6	16.69±0.9	41.26±1.8	44.37±1.2	46.52±1.4
PFEDMXF	56.99±2.8	61.38±2.1	65.56±2.1	45.07±1.5	46.88±1.4	47.78±1.5	13.05±0.7	15.27±0.7	16.58±0.8	40.74±1.3	43.92±1.7	45.98±1.5
FEDFMX(Ours)	**60.35**±2.6	**65.28**±1.8	**68.94**±2.1	**49.87**±1.2	**51.18**±1.3	**53.24**±1.2	**16.76**±0.8	**19.47**±0.5	**21.36**±0.6	**44.82**±1.5	**47.63**±1.1	**49.76**±1.2

(b) Under the DIL setting

Method	CIFAR-10			CIFAR-100			Tiny-ImageNet			DomainNet		
	$\alpha_D=0.1$	$\alpha_D=0.5$	$\alpha_D=1.0$	$\alpha_D=0.1$	$\alpha_D=0.5$	$\alpha_D=1.0$	$\alpha_D=0.1$	$\alpha_D=0.5$	$\alpha_D=1.0$	$\alpha_D=0.1$	$\alpha_D=0.5$	$\alpha_D=1.0$
FEDMUT	15.62±2.3	20.11±2.1	26.48±2.7	4.21±1.3	6.37±1.2	8.54±1.0	2.41±0.8	3.58±0.6	4.96±0.7	3.80±1.4	5.61±1.3	7.72±1.0
FEDCROSS	17.38±2.6	22.74±2.1	28.93±2.2	5.03±0.9	7.41±1.3	9.86±1.4	2.86±0.6	4.21±0.7	5.64±0.7	4.46±1.2	6.52±0.9	8.68±1.1
FEDLSA	13.47±2.0	18.23±2.0	23.94±2.1	3.56±1.1	5.48±0.9	7.63±1.1	2.02±0.5	3.01±0.6	4.27±0.8	3.13±0.8	4.82±0.9	6.71±1.0
FEDWMSAM	16.82±1.8	21.93±2.1	27.41±2.2	4.76±1.8	6.89±0.9	9.12±1.5	2.63±0.9	3.89±0.6	5.31±0.7	4.19±1.2	6.06±1.4	8.03±1.1
PFEDMOE	39.84±2.4	46.27±1.9	52.18±2.1	22.61±1.4	28.73±1.3	33.24±1.4	7.92±0.8	10.84±0.8	12.63±0.9	19.92±1.4	25.28±1.8	29.25±1.3
FEDEWC	16.24±2.8	20.07±2.6	29.44±2.3	8.22±1.0	11.03±0.8	12.31±1.6	3.62±0.5	4.39±0.6	5.22±0.8	7.23±1.7	9.71±1.3	10.83±1.6
FEDLWF	34.82±3.1	38.98±2.3	51.53±2.8	18.96±1.1	25.21±0.7	28.97±1.2	3.97±0.4	6.25±0.9	10.34±1.1	16.68±1.3	22.18±1.9	25.49±1.4
TARGET	32.21±2.2	35.76±1.7	48.62±1.8	14.04±0.9	21.38±0.7	25.35±1.3	5.21±0.6	5.67±0.8	7.86±0.7	12.36±1.5	18.81±0.9	22.31±1.2
LANDER	35.73±2.8	39.57±2.3	58.19±2.5	27.35±1.8	34.02±1.5	37.66±1.8	8.78±0.7	8.66±0.5	10.24±0.6	24.07±1.7	29.94±1.2	33.14±1.4
RE-FED	51.02±3.4	54.21±3.2	58.36±2.3	29.42±1.4	39.76±1.2	40.87±1.2	9.18±0.8	11.43±0.7	12.28±0.9	25.89±1.5	34.99±1.7	35.97±1.1
FL-CLIP	49.37±3.2	52.19±2.8	57.85±2.4	28.59±1.5	37.95±1.4	39.90±1.4	8.94±0.7	10.91±0.8	12.34±1.0	25.16±1.4	33.40±1.2	35.11±1.3
FEDCBDR	56.49±2.4	60.58±1.6	64.96±1.9	44.63±1.3	46.27±1.4	47.42±1.3	12.29±0.6	14.38±0.6	15.72±0.7	39.27±1.1	42.72±1.5	43.73±1.6
PFEDMXF	55.78±2.6	60.43±1.5	64.82±1.9	43.97±1.4	46.09±1.6	47.12±1.3	12.17±0.5	14.25±0.7	15.46±0.6	38.69±1.2	40.56±1.0	41.47±1.1
FEDFMX(Ours)	**59.36**±2.1	**64.68**±1.2	**67.46**±1.5	**47.87**±1.1	**50.32**±1.5	**50.76**±1.4	**15.88**±0.5	**19.14**±0.7	**22.95**±0.4	**43.75**±0.9	**44.29**±1.1	**46.83**±1.05

where the ViT-B/16 model is initialized with self-supervised pretrained weights from DINO [2], which is widely used in CIL.

Baselines. We compare our method with following three types of baselines: (i) General FL methods: FEDMUT [19], FEDCROSS [20], FEDLSA [12], FEDWMSAM [30], PFEDMOE [64]; (ii) Traditional CIL methods under federated settings: FEDEWC [24], FEDLWF [37]; (iii) FCIL methods: TARGET [70], LANDER [58], RE-FED [32], FL-CLIP [56], FEDCBDR [49], PFEDMXF [71].

Implementation Details. We employ SGD as the optimizer, a learning rate of 0.01, momentum of 0.9, and weight decay of $1e^{-5}$. For all datasets, we set the batch size to 64 and the local epoch to 5. We adopt two different types of Non-IID settings [28]: Quantity-based Label Imbalance (QLI) and Distribution-based Label Imbalance (DLI), denoted by α_Q and α_D respectively, where BLI allocates the data samples via the Dirichlet distribution. To assess robustness under varying task granularity, we consider the following two incremental settings: *coarse-grained* (CIFAR-10 divided into 3 tasks, CIFAR-100 into 5 tasks, DomainNet into 5 tasks, and Tiny-ImageNet into 10 tasks); *fine-grained* (CIFAR-10 divided into 5 tasks, CIFAR-100 into 10 tasks, DomainNet into 10 tasks, and Tiny-ImageNet into 20 tasks). The weight ρ for (2) is set to 0.95, and $\alpha = 0.6$ for (5). The regularization terms in (10) is weighted by $\beta = 0.05$ and $\gamma = 0.6$. Unless otherwise specified, we set $\alpha_D = 0.5$ by default. We set $N = 100$ clients, where 20% of clients are randomly sampled at each round. We default the num-

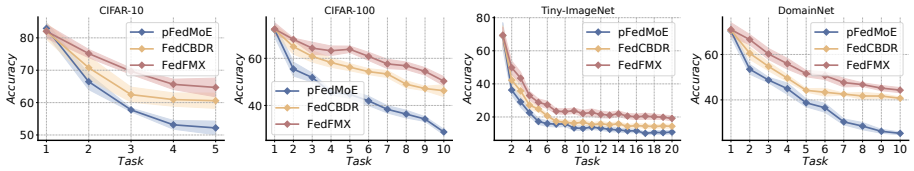

Fig. 2: Performance comparison under the fine-grained setting.

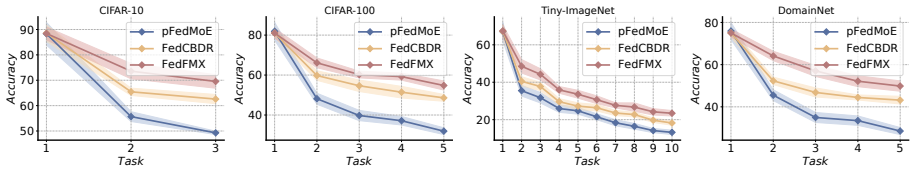

Fig. 3: Performance comparison under the coarse-grained setting.

Table 2: Performance comparisons of expert selection strategies during training.

Non-IID Settings	Fine-Grained Increment									Coarse-Grained Increment								
	CIFAR-100			Tiny-ImageNet			DomainNet			CIFAR-100			Tiny-ImageNet			DomainNet		
	Top-1	Top-K	AES	Top-1	Top-K	AES	Top-1	Top-K	AES	Top-1	Top-K	AES	Top-1	Top-K	AES	Top-1	Top-K	AES
QLI	46.38	49.12	**51.18**	15.83	17.72	**19.47**	42.38	45.92	**47.63**	51.03	53.45	**55.63**	20.63	22.81	**24.06**	46.71	48.95	**51.84**
DLI	45.96	48.54	**50.32**	15.17	16.88	**19.14**	41.59	44.86	**44.29**	50.76	52.94	**54.78**	19.75	20.87	**23.42**	45.89	48.12	**49.85**

ber of experts to $K = 4$, where each expert shares the same architecture but maintains independent parameters, consisting of an expert body and a classification head built on top of the shared backbone. During inference, we select the Top-2 experts according to the routing scores. We conduct experiments on Ubuntu 22.04.4; GPU: 2NVIDIA GeForce RTX 4090. More experiment details are provided in Appendix F.

5.2 Main Results and Analysis

We evaluate the performance of FedFMX under two label imbalance scenarios, namely QLI and DLI, as summarized in Table 1a and Table 1b. Our method consistently achieves superior performance compared with benchmark methods, demonstrating the effectiveness of FedFMX. Under the QLI setting, FedFMX maintains stable performance as the degree of label imbalance increases, whereas conventional approaches tend to suffer from degraded generalization due to intensified forgetting and capacity conflicts, highlighting the advantage of the proposed Fisher-guided routing and adaptive expert selection mechanisms, which effectively coordinate expert specialization while mitigating gradient inconsistencies caused by skewed class distributions. The advantage of FedFMX becomes more evident on challenging datasets such as Tiny-ImageNet and DomainNet, where the number of classes and data complexity are substantially higher. In such scenarios, the proposed framework remains robust and continues to outperform competing approaches, indicating its strong capability to generalize to high-dimensional and heterogeneous federated environments.

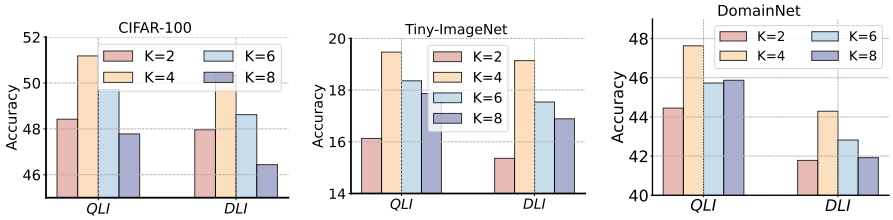

Fig. 4: Performance on the quantity of total experts under the fine-grained setting.

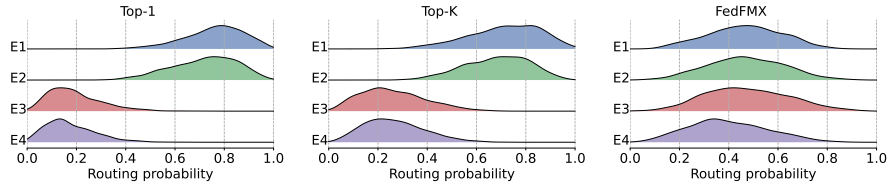

Fig. 5: Comparison of expert routing probability distributions on DomainNet.

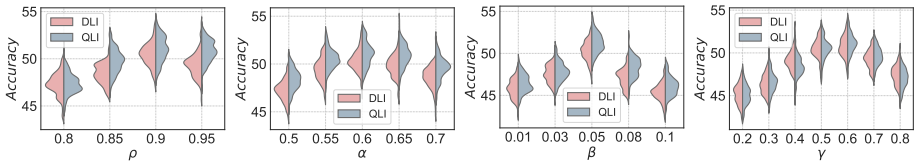

Fig. 6: Hyperparameters sensitivity on CIFAR-100 under the QLI/DLI setting.

5.3 Performance under Incremental Tasks

Figures 2-3 illustrate the performance evolution across incremental tasks under the fine- and coarse-grained settings, where $\alpha_D = 0.5$ by default. As new tasks are introduced, the performance of all methods gradually declines due to catastrophic forgetting. Nevertheless, FedFMX consistently maintains higher accuracy throughout the incremental process. The performance gap between FedFMX and the baseline methods becomes increasingly evident as the number of tasks grows. While competing approaches exhibit noticeable degradation in later stages, FedFMX shows a significantly slower decline, indicating that the proposed Fisher-guided routing and adaptive expert selection strategies effectively preserve previously acquired knowledge while still enabling the model to adapt to newly introduced classes, demonstrating stronger resilience to drastic task transitions and improved stability over long incremental sequences.

5.4 Expert Quantity Study

To validate the effectiveness of our AES module, we compare with Top-1 and Top-K routing strategies under both QLI/DLI and fine-/coarse-grained incremental settings. As depicted in Table 2, AES consistently achieves the best performance across all datasets and settings. Compared with Top-1 routing, which

Table 3: Efficacy of each module on various datasets.

(a) Under the QLI setting

Module			Fine-Grained Increment			Coarse-Grained Increment		
FRES	AES	RAR	CIFAR-100	Tiny-ImageNet	DomainNet	CIFAR-100	Tiny-ImageNet	DomainNet
			33.24	10.86	29.41	38.57	13.39	34.18
	✔	✔	48.33	17.96	45.58	53.41	22.51	49.92
✔		✔	42.71	15.84	40.83	48.26	19.87	45.62
✔	✔		46.23	16.21	44.12	51.26	21.63	48.67
✔	✔	✔	**51.18**	**19.74**	**47.63**	**55.63**	**24.06**	**51.84**

(b) Under the DIL setting

Module			Fine-Grained Increment			Coarse-Grained Increment		
FRES	AES	RAR	CIFAR-100	Tiny-ImageNet	DomainNet	CIFAR-100	Tiny-ImageNet	DomainNet
			32.11	10.42	26.83	37.69	12.94	31.88
	✔	✔	47.98	17.41	42.71	52.89	21.88	47.83
✔		✔	41.86	15.32	38.34	47.63	19.46	44.02
✔	✔		45.71	16.76	41.63	50.91	21.17	47.12
✔	✔	✔	**50.32**	**19.14**	**44.29**	**54.78**	**23.42**	**49.85**

activates only a single expert, AES provides improved performance by enabling multiple experts to collaboratively capture diverse task-specific representations. Meanwhile, although Top-K routing increases the number of activated experts, its fixed selection strategy cannot adapt to the varying importance of experts across different tasks and data distributions. In contrast, AES dynamically determines the expert subset based on the learned importance signals, allowing the model to better balance expert specialization and knowledge sharing.

To investigate the influence of expert quantity on model performance, we vary the expert pool size $K \in \{2, 4, 6, 8\}$ under both QLI and DLI scenarios with $\alpha_D = 0.5$ and the fine-grained incremental task setting, as shown in Figure 4. The results indicate that increasing the quantity of experts initially improves the performance, suggesting that a moderate expansion of expert capacity enables the capture of diverse task-specific knowledge. However, with an excessively large expert pool, performance decreases, implying that although a larger number of experts increases representational capacity, it may also introduce stronger routing competition and knowledge fragmentation, thereby weakening the effectiveness of expert specialization. Consequently, the moderate quantity of experts achieves a better balance between model capacity and expert coordination.

To further study the expert utilization under different routing strategies, in Figure 5, we analyze the routing probability distributions of individual experts on DomainNet under Top-1/Top-K routing, and FedFMX. Top-1 routing leads to highly skewed expert utilization, where only a small subset of experts receives high routing probabilities while others are rarely activated, limiting expert diversity and leading to capacity imbalance. Top-K routing partially alleviates this issue, but the routing probabilities remain biased, resulting in uneven expert participation. In contrast, FedFMX produces a noticeably more balanced routing distribution, indicating that different experts are utilized more consistently,

Table 4: Computational cost analysis under different backbone architectures.

(a) ResNet-18

Method	CIFAR-100				Tiny-ImageNet				DomainNet			
	Params	FLOPs	Time(s)	Acc	Params	FLOPs	Time(s)	Acc	Params	FLOPs	Time(s)	Acc
pFedMoE	11.7M	89	75	30.76	11.9M	487	375	11.02	12.0M	1573	826	27.92
FedFMX	14.6M	98	82	51.18	14.7M	588	412	19.47	15.0M	1785	934	47.63

(b) ViT-B/16

Method	CIFAR-100				Tiny-ImageNet				DomainNet			
	Params	FLOPs	Time(s)	Acc	Params	FLOPs	Time(s)	Acc	Params	FLOPs	Time(s)	Acc
pFedMoE	89.5M	902	645	45.37	89.7M	928	647	22.51	89.8M	928	692	41.25
FedFMX	93.0M	974	696	68.44	93.2M	984	716	30.92	93.3M	1015	759	63.94

which suggests that our method effectively promotes expert diversity while mitigating expert collapse, and better exploits the capacity of the expert pool.

5.5 Hyerparameter Sensitivity Analysis

We evaluate the robustness of our method regarding hyperparameter choices and conduct a sensitivity analysis on CIFAR-100 under the DLI/QLI setting by varying four key parameters. As illustrated in Fig. 6, FedFMX maintains relatively stable performance across a wide range of parameter values. Specifically, for ρ, the performance first improves and then slightly decreases as ρ increases, indicating that a moderate sparsity level provides a better balance between model capacity and structural regularization. Similar trends can be observed for α and β. The results demonstrate that FedFMX achieves robust performance without requiring delicate hyperparameter tuning, highlighting the stability and practical applicability of the proposed framework.

5.6 Ablation Studies

In Table 3, we evaluate the contribution of each component in FedFMX and conduct an ablation study by progressively enabling the three key modules. Each module consistently improves the overall performance across different datasets and incremental settings. Removing any component leads to noticeable performance degradation, indicating complementary benefits. In particular, FRES improves expert routing quality by guiding the model to identify stable and informative experts, while AES further enhances the routing process by adaptively selecting expert subsets based on their estimated importance. Meanwhile, RAR introduces additional regularization that encourages more balanced expert utilization and stabilizes training during incremental updates. When all components are jointly enabled, the model achieves the best performance across both QLI and DLI settings as well as across fine-grained and coarse-grained increments. These results demonstrate that the three modules cooperate effectively to improve expert routing, mitigate knowledge interference, and enhance robustness.

5.7 Communication Studies

To evaluate the computational efficiency, we compare FedFMX with pFedMoE under two backbone architectures under the fine-grained incremental setting. As shown in Table 4, FedFMX introduces only a modest increase in computational cost compared with pFedMoE. Despite this overhead, FedFMX consistently achieves significantly better model performance, which is particularly notable on challenging datasets, such as Tiny-ImageNet and DomainNet. These results indicate that the proposed routing and expert selection mechanisms effectively improve model capacity and knowledge retention while maintaining a reasonable computational overhead, demonstrating a favorable trade-off between efficiency and performance.

6 Conclusion

In this paper, we presented FedFMX, a Mixture-of-Experts-based framework tailored for FCIL. To address the fundamental challenges of catastrophic forgetting, data heterogeneity, and synchronized class misalignment, FedFMX introduces a Fisher-Routed Expert Scoring module that quantifies each expert's stability–plasticity trade-off, an Adaptive Expert Selection module based on marginal contributions to dynamically activate expert subsets, and a routing-aware regularization scheme that balances expert utilization and supports efficient inference. By unifying these components, FedFMX enables stable incremental learning, expert specialization, and flexible deployment without relying on ground-truth labels during inference. Rigorous theoretical analyses prove the $\mathcal{O}(T^{-1})$ convergence rate. Experiments across multiple benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches, validating the effectiveness for realistic and scalable FCIL scenarios.

Acknowledgements

This work was supported by the UGC General Research Fund no. 17209822 and the Innovation and Technology Commission Fund no. ITS/383/23FP from Hong Kong.

References

1. Babakniya, S., Fabian, Z., He, C., Soltanolkotabi, M., Avestimehr, S.: A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. Advances in Neural Information Processing Systems **36**, 66408–66425 (2023)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

3. Chen, X., Zhang, B., Zhou, X., Sun, M., Zhang, S., Zhang, S., Li, G.Y.: Efficient training of large-scale ai models through federated mixture-of-experts: A system-level approach. arXiv preprint arXiv:2507.05685 (2025)
4. Chi, Z., Dong, L., Huang, S., Dai, D., Ma, S., Patra, B., Singhal, S., Bajaj, P., Song, X., Mao, X.L., et al.: On the representation collapse of sparse mixture of experts. Advances in Neural Information Processing Systems **35**, 34600–34613 (2022)
5. Dong, J., Li, H., Cong, Y., Sun, G., Zhang, Y., Van Gool, L.: No one left behind: Real-world federated class-incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2054–2070 (2023)
6. Dong, J., Wang, L., Fang, Z., Sun, G., Xu, S., Wang, X., Zhu, Q.: Federated class-incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10164–10173 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Fan, T., Gu, H., Cao, X., Chan, C.S., Chen, Q., Chen, Y., Feng, Y., Gu, Y., Geng, J., Luo, B., et al.: Ten challenging problems in federated foundation models. IEEE Transactions on Knowledge and Data Engineering (2025)
9. Farhat, Y., Shili, H.E., Liao, F., Dun, C., Garcia, M.D.C.H., Zheng, G., Awadallah, A.H., Sim, R., Dimitriadis, D., Kyrillidis, A.: Learning to specialize: Joint gating-expert training for adaptive moes in decentralized settings. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
10. Feng, J., Yang, X., Liang, L., Han, W., Fang, B., Liao, Q.: Cgofed: Constrained gradient optimization strategy for federated class incremental learning. IEEE Transactions on Knowledge and Data Engineering (2025)
11. Feng, Y., Geng, Y.a., Zhu, Y., Han, Z., Yu, X., Xue, K., Luo, H., Sun, M., Zhang, G., Song, M.: Pm-moe: Mixture of experts on private model parameters for personalized federated learning. In: Proceedings of the ACM on Web Conference 2025. pp. 134–146 (2025)
12. Fu, L., Huang, S., Lai, Y., Liao, T., Zhang, C., Chen, C.: Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 16648–16656 (2025)
13. Gao, X., Yang, X., Yu, H., Kang, Y., Li, T.: Fedprok: Trustworthy federated class-incremental learning via prototypical feature knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4205–4214 (2024)
14. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3614–3631 (2020)
15. Guan, Z., Zhu, G., Zhou, Y., Liu, W., Wang, W., Luo, J., Gu, X.: Stsa: Federated class-incremental learning via spatial-temporal statistics aggregation. arXiv preprint arXiv:2506.01327 (2025)
16. Guo, H., Zhu, F., Liu, W., Zhang, X.Y., Liu, C.L.: Pilora: Prototype guided incremental lora for federated class-incremental learning. In: European Conference on Computer Vision. pp. 141–159. Springer (2024)
17. He, J., Duan, Z., Zhu, F.: Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30534–30544 (2025)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hu, M., Cao, Y., Li, A., Li, Z., Liu, C., Li, T., Chen, M., Liu, Y.: Fedmut: Generalized federated learning via stochastic mutation. In: Proceedings of the AAAI conference on artificial intelligence. pp. 12528–12537 (2024)
20. Hu, M., Zhou, P., Yue, Z., Ling, Z., Huang, Y., Li, A., Liu, Y., Lian, X., Chen, M.: Fedcross: Towards accurate federated learning via multi-model cross-aggregation. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE). pp. 2137–2150. IEEE (2024)
21. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al.: Advances and open problems in federated learning. Foundations and trends® in machine learning **14**(1–2), 1–210 (2021)
22. Karimireddy, S.P., Rebjock, Q., Stich, S., Jaggi, M.: Error feedback fixes SignSGD and other gradient compression schemes. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3252–3261. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/karimireddy19a.html
23. Ke, H., Shi, J., Zhang, Y., Wang, F., Xie, Y., Qu, Y.: Task-aware prompt gradient projection for parameter-efficient tuning federated class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2631–2641 (2025)
24. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
25. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
26. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
27. Li, D., Zeng, Z., Dai, W., Suganthan, P.N.: Complementary learning subnetworks towards parameter-efficient class-incremental learning. IEEE Transactions on Knowledge and Data Engineering (2025)
28. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: 2022 IEEE 38th international conference on data engineering (ICDE). pp. 965–978. IEEE (2022)
29. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine **37**(3), 50–60 (2020)
30. Li, T., Huang, Y., Jiang, L., Liu, C., Xie, Q., Du, W., Wang, L., Wu, K.: FedWMSAM: Fast and flat federated learning via weighted momentum and sharpness-aware minimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=75JiIa0fU1
31. Li, Y., Liu, Y., Guo, B., Wang, D., Li, H., Li, N., Wang, Y., Luo, H., Yu, Z.: Cross-f ^{2} scil: A federated few-shot class incremental learning method for cross mobile edge network environments. IEEE Transactions on Services Computing (2025)
32. Li, Y., Li, Q., Wang, H., Li, R., Zhong, W., Zhang, G.: Towards efficient replay in federated incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12820–12829 (2024)

33. Li, Y., Wang, H., Qi, Y., Liu, W., Li, R.: Re-fed+: A better replay strategy for federated incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
34. Li, Y., Wang, Y., Dong, J., Wang, H., Qi, Y., Zhang, R., Li, R.: Resource-constrained federated continual learning: What does matter? In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
35. Li, Y., Wang, Y., Wang, H., Qi, Y., Xiao, T., Li, R.: FedSSI: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In: Forty-second International Conference on Machine Learning (2025)
36. Li, Y., Xu, W., Qi, Y., Wang, H., Li, R., Guo, S.: Sr-fdil: Synergistic replay for federated domain-incremental learning. IEEE Transactions on Parallel and Distributed Systems **35**(11), 1879–1890 (2024)
37. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
38. Liang, F.Y., Zhan, Y.W., Liu, J., Zhang, C.Y., Chen, Z.D., Luo, X., Xu, X.S.: Class-aware prompting for federated few-shot class-incremental learning. IEEE Transactions on Circuits and Systems for Video Technology (2025)
39. Liang, T., Hu, M., Sun, E.: Mixture of specialized experts for model-heterogeneous personalized federated learning. IEEE Networking Letters (2025)
40. Liu, Y., Huang, D.: Sparse personalized federated class-incremental learning. Information Sciences **706**, 121992 (2025)
41. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**(2), 275–293 (2014)
42. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
43. Mei, H., Cai, D., Zhou, A., Wang, S., Xu, M.: Fedmoe: Personalized federated learning via heterogeneous mixture of experts. arXiv preprint arXiv:2408.11304 (2024)
44. Miao, C., Chang, T., Wu, M., Xu, H., Li, C., Li, M., Wang, X.: Fedvla: Federated vision-language-action learning with dual gating mixture-of-experts for robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6904–6913 (2025)
45. Mu, S., Lin, S.: A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. arXiv preprint arXiv:2503.07137 (2025)
46. Nori, M.K., KIM, I.M., Wang, G.: Autoencoder-based hybrid replay for class-incremental learning. In: Forty-second International Conference on Machine Learning (ICML) (2025)
47. Nori, M.K., KIM, I.M., Wang, G.: Federated class-incremental learning: A hybrid approach using latent exemplars and data-free techniques to address local and global forgetting. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025)
48. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)
49. Qi, Z., Tang, Y.P., Meng, L., Yu, H., Li, X., Meng, X.: Class-wise balancing data replay for federated class-incremental learning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
50. Qiu, Z., Huang, Z., Zheng, B., Wen, K., Wang, Z., Men, R., Titov, I., Liu, D., Zhou, J., Lin, J.: Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. arXiv preprint arXiv:2501.11873 (2025)

51. Radwan, A., Soliman, M., Abdelaziz, O., Shehata, M.: Feddg-moe: Test-time mixture-of-experts fusion for federated domain generalization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1811–1820 (2025)
52. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
53. Shi, Y., Tong, Y., Zeng, Y., Zhou, Z., Ding, B., Chen, L.: Efficient approximate range aggregation over large-scale spatial data federation. IEEE Transactions on Knowledge and Data Engineering **35**(1), 418–430 (2021)
54. Stich, S.U., Cordonnier, J.B., Jaggi, M.: Sparsified sgd with memory. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://proceedings.neurips.cc/paper_files/paper/2018/file/b440509a0106086a67bc2ea9df0a1dab-Paper.pdf
55. Sun, R., Duan, H., Dong, J., Ojha, V., Shah, T., Ranjan, R.: Rehearsal-free federated domain-incremental learning. In: 2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS). pp. 835–845. IEEE (2025)
56. Tan, A.Z., Feng, S., Yu, H.: Fl-clip: Bridging plasticity and stability in pre-trained federated class-incremental learning models. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
57. Tan, C.H., Chen, Q., Wang, W., Ma, Y., Zhang, C., Deng, C., Zhang, Q., Li, X., Ye, J.: Fggm: Fisher-guided gradient masking for continual learning (2026)
58. Tran, M.T., Le, T., Le, X.M., Harandi, M., Phung, D.: Text-enhanced data-free approach for federated class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23870–23880 (2024)
59. Wang, Q., Chen, S., Wu, M., Li, X.: Digital twin-empowered federated incremental learning for non-iid privacy data. IEEE Transactions on Mobile Computing (2024)
60. Wu, F., Feng, S., Chen, Y., Zhao, L.: Personalized federated class-incremental learning through critical parameter transfer. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
61. Xie, L., Luan, T., Cai, W., Yan, G., Chen, Z., Xi, N., Fang, Y., Shen, Q., Wu, Z., Yuan, J.: dflmoe: Decentralized federated learning via mixture of experts for medical data analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10203–10213 (2025)
62. Yang, X., Yu, H., Gao, X., Wang, H., Zhang, J., Li, T.: Federated continual learning via knowledge fusion: A survey. IEEE Transactions on Knowledge and Data Engineering **36**(8), 3832–3850 (2024)
63. Yi, L., Yu, H., Ren, C., Zhang, H., Wang, G., Liu, X., Li, X.: pfedmoe: Data-level personalization with mixture of experts for model-heterogeneous personalized federated learning. arXiv preprint arXiv:2402.01350 (2024)
64. Yi, L., Yu, H., Wang, G., Liu, X., Hu, Q.: pfedmoe: Data-level personalization with mixture of experts in model-heterogeneous personalized federated learning. IEEE Transactions on Knowledge and Data Engineering **38**(3), 1905–1918 (2026). https://doi.org/10.1109/TKDE.2026.3656194
65. You, Z., Chu, J., Li, Z., Liu, B., Li, T.: Adaptive federated class-incremental learning for reducing catastrophic forgetting. Expert Systems with Applications p. 128442 (2025)
66. Yu, F., Hu, J., Min, G.: Efficient federated class-incremental learning of pre-trained models via task-agnostic low-rank residual adaptation. arXiv preprint arXiv:2505.12318 (2025)

67. Yu, H., Yang, X., Gao, X., Feng, Y., Wang, H., Kang, Y., Li, T.: Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 5280–5288 (2024)
68. Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty years of mixture of experts. IEEE transactions on neural networks and learning systems **23**(8), 1177–1193 (2012)
69. Zhan, Z., Zhao, W., Li, Y., Liu, W., Zhang, X., Tan, C.W., Wu, C., Guo, D., Chen, X.: Fedmoe-da: Federated mixture of experts via domain aware fine-grained aggregation. In: 2024 20th International Conference on Mobility, Sensing and Networking (MSN). pp. 122–129. IEEE (2024)
70. Zhang, J., Chen, C., Zhuang, W., Lyu, L.: Target: Federated class-continual learning via exemplar-free distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4782–4793 (2023)
71. Zhang, Y., Zhu, H., Tan, A.Z., Yu, D., Huang, L., Yu, H.: pfedmxf: Personalized federated class-incremental learning with mixture of frequency aggregation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30640–30650 (2025)
72. Zhong, Z., Bao, W., Wang, J., Chen, J., Lyu, L., Lim, W.Y.B.: Sacfl: Self-adaptive federated continual learning for resource-constrained end devices. IEEE Transactions on Neural Networks and Learning Systems (2025)
73. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)
74. Zhuang, Y., Li, Y., Song, Y., Qiu, M.: Personalized federated learning for fault diagnosis with mixture of experts. Information Fusion p. 103439 (2025)