Supplemental Document for Federated Few-Shot Class-Incremental Learning

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This document presents supplemental document for Federated Few-Shot Class-Incremental Learning that includes the detailed algorithm of UOPP that is presented in section 1, the detailed theoretical analysis presented in section 2, the detailed experimental setting that is presented in section 3, the detailed numerical results that are presented in section 4, the detailed results on stability-plasticity that are presented in section 5, detailed numerical results on variation of local clients and global rounds that are presented in section 6, detailed numerical results on ablation study that are presented in section 7, and detailed complexity analysis that is presented in section 8.**

***Index Terms*—Federated, Few-Shot, Class-Incremental Learning**

I. Detailed Process of Unified Optimized Prototype Prompt (UOPP)

In this section, we present the detailed algorithm of UOPP as shown in algorithm 1.

II. Detailed Theoretical Analysis

Let $\Theta = (P, \Phi, \Psi)$ be the trainable parameters, $F(\Theta) = \mathbb{E}[\mathcal{L}(\mathcal{T}; \Theta)] = \mathbb{E}[\mathcal{L}(\mathcal{T}; (P, \Phi, \Psi))]$ is the expected loss function, k, E, R, and L is local iteration, local epoch, global round, and number of selected local clients respectively. Please note that in this analysis, L denotes the number of selected local clients, while $l \geq 1$ denotes a constant for the l-smooth coefficient. Following the update rule in section 4.3, the expression of $F(\Theta)$ above can be detailed as follows:

(i) Base Task ($t = 0$): $\Theta = (P, \Phi)$, and $F(\Theta) = \mathbb{E}[\mathcal{L}(\mathcal{T}; \Theta)] = \mathbb{E}[\mathcal{L}_{l+}(\mathcal{T}; (P, \Phi))]$ as local clients update (P, Φ) using $\mathcal{L}_{l+}$ following equations 7 and 8.

(i) FS Task ($t \geq 1$): $\Theta = (P, \Psi)$, and $F(\Theta) = \mathbb{E}[\mathcal{L}(\mathcal{T}; \Theta)] = \mathbb{E}[\mathcal{L}_{lfs+}(\mathcal{T}; (P, \Psi))]$ as local clients update (P, Ψ) using $\mathcal{L}_{lfs+}$ following equations 9 and 10.

We adopt the SGD optimization convergence analysis [1] and FedAvg convergence analysis [2] assumptions as follows:

Assumption 1: $F_1, ...F_l, ..., F_{L_S}$ are all $L-$smooth: for all Θ and Θ', $F_l(\Theta) \leq F_l(\Theta') + (\Theta - \Theta')^T \nabla F_l(\Theta) + \frac{L}{2}||\Theta - \Theta'||_2^2$.

Assumption 2: $F_1, ...F_l, ..., F_{L_S}$ are all $\mu-$strongly convex: for all Θ and Θ', $F_l(\Theta) \leq F_l(\Theta') + (\Theta - \Theta')^T \nabla F_l(\Theta) + \frac{\mu}{2}||\Theta - \Theta'||_2^2$.

Assumption 3: Let ξ_l^k be the random uniformly sampled from l-th local data at $k - th$ iteration . The variance of stochastic gradients in each client is bounded by the following criteria: $\mathbb{E}||\nabla F_l(\Theta_l^k, \xi_l^k) - \nabla F_l(\Theta_l^k)|| \leq \sigma_l^2$ for $l = 1, 2, ..., L_S$

Assumption 4: The expected squared norm of stochastic gradients in each client is bounded by: $\mathbb{E}||\nabla F_l(\Theta_l^k, \xi_l^k)|| \leq G^2$ for all $l = 1, 2, ..., L_S$ and $k = 1, 2, ..., K$ where $K \in \mathbb{N}$.

Assumption 5: $\sum_{k=1}^{\infty} \alpha_l^k = \infty$ and $\sum_{k=1}^{\infty} {\alpha_l^k}^2 < \infty$ where α_l^k is the learning rate of $l - th$ client in k-th step training.

A. *Proof of Theorem 1*

Let a client S_l be trained locally with its local data $\mathcal{T}_l^t \cup Z^t$, where $\mathcal{T}_l^t$ is local;y observed training samples for t-th task and $Z^t = Z_G^t$ is aggregated unified prototype for task t shared by server respectively. We assume that Z^t is augmented so that $|z_{c_b}^t| \approx |x_{c_a}|$ for $z_{c_b}^t \in Z^t$ and $x_{c_a}^t \in \mathcal{T}_l^t \subseteq \mathcal{T}^t$. Ias the implication, the number of prototypes of unavailable classes in $\mathcal{T}_l^t$ and the samples of available classes in $\mathcal{T}_l^t$ are balanced. Then the local model $\Theta_l = (P_l, \Phi_l)$ or $\Theta_l = (P_l, \Psi_l)$ is updated in K iterations based on minibatches drawn from $\mathcal{T}_l^t \cup Z^t$. Since the backbone (feature extractor) is frozen, and $\mathcal{T}_l^t \cup Z^t$ has balance samples for all classes, then ξ_l^k approximates ξ^k that is a sample from $\mathcal{T}^t$. The local model is updated by the stochastic gradient (SG) approach as presented in equations (6) and (10) in the main paper. Suppose that $g(\Theta_l, \xi_l^k)$ is the stochastic gradient function, then the update process can be simplified as:

$$\Theta_l^{k+1} \leftarrow \Theta_l^k - \alpha_l^k g(\Theta_l^k, \xi_l^k) \tag{A1}$$

Under assumption 1, and local training updates Θ by iterating SG with sample ξ_l^k, then we have:

$$\begin{aligned} F_l(\Theta_l^{k+1}) - F_l(\Theta_l^k) &\leq (\Theta_l^{k+1} - \Theta_l^k)^T \nabla F_l(\Theta_l^k) + \frac{L}{2}||\Theta_l^{k+1} - \Theta_l^k||_2^2 \\ &\leq -\alpha_l^k \nabla F_l(\Theta_l^k)^T g(\Theta_l^k, \xi_l^k) + {\alpha_l^k}^2 \frac{L}{2}||g(\Theta_l^k, \xi_l^k)||_2^2 \end{aligned} \tag{A2}$$

The equation above can be derived into:

$$\begin{aligned} \mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - F_l(\Theta_l^k) \leq &- \alpha_l^k \nabla F_l(\Theta_l^k)^T \mathbb{E}[g(\Theta_l^k, \xi_l^k)] \\ &+ {\alpha_l^k}^2 \frac{L}{2} \mathbb{E}_{\xi_l^k}[||g(\Theta_l^k, \xi_l^k)||_2^2] \end{aligned} \tag{A3}$$

The inequation above shows Θ_l^k optimization by SG method at a step k, and it shows the reduction of F_l (left side) is bounded by a quantity in the right side involving ∇F_l which is directional derivative of F_l at Θ_l^k along with $-g(\Theta_l^k, \xi_l^k)$ (first term) and second moment of $g(\Theta_l^k, \xi_l^k)$ (second term).

Algorithm 1 UOPP

Input: Number of clients N, number of selected local clients L, total number of rounds R, number of task $T+1$, local epochs E, batch size B.
Distribute frozen ViT backbone f to all clients $\{S_l\}_{l=1}^{N}$ and central server S_G
Initiate prompt, key, and head layer for all clients and central server $P_G = P_l,\ \Phi_G = \Phi_l,\ \Psi_l = init(), l \in \{1..N\}$
$R_T \leftarrow R/(T+1)$, R_T represents round per task
Init global and local unified prototypes $Z_G^t = Z_l^t, = Z^t = \emptyset$
for $t = 0 : T$ **do**
  for $r = 1 : R_T$ **do**
    $S_l \leftarrow$ randomly select L local clients from N total clients
    Clients execute:
    if $R_T = 1$ **then**
      Compute static prototype $\tilde{Z}_l^t$ as in Eq. (1) to (5), then send it to server
    end if
    Receive global parameters i.e. prompt, FC layer, and prototypes set P_G, Φ_G, and Z_G^t
    Assign local parameters $(P_l, \Phi_l, Z_l^t) \leftarrow (P_G, \Phi_G, Z_G^t)$
    $\mathcal{B} \leftarrow$ Split $\mathcal{T}_l^t$ into B sized batches
    for $e = 1 : E$ **do**
      for $b = 1 : \mathcal{B}$ **do**
        if $(t = 0)$ **then** // Base Task Update
          Compute prompt-generated feature $f_{P_l}(x)$ as in Eq. (1) to (3)
          Compute logits with FC clsasifier $g_{\Phi_l}(f_{P_l}(x) \cup Z_G^t)$
          Compute loss $\mathcal{L}_{l+}$ as in Eq. (7)
          Update local parameters (P_l, Φ_l) based on $\mathcal{L}_{t+}$ as in Eq. (8)
        else $(t \geq 1)$ // Few-shot Task Update
          Compute static prototype $\tilde{Z}_l^t$ using feature $f_{P_l}(x)$ as in Eq. (1) to (5)
          Draw $\mathcal{S}$ from $\tilde{Z}_l^t$ and draw $\mathcal{Q}$ from $Z_l^t = Z_G^t$
          Rectify dynamic prototype $\hat{Z}_l^t$ using $g_\Psi(.)$as in Eq. (11) to (14)
          Form unified prototype $Z_l^t = Z_G^t \cup \hat{Z}_l^t$
          Compute logits with PB classifier $g_{Z_l^t}(f_{P_l}(x) \cup \mathcal{S})$
          Compute loss $\mathcal{L}_{lfs+}$ as in Eq. (9)
          Update local parameters (P_l, Ψ_l) based on $\mathcal{L}_{lfs+}$ as in Eq. (10)
        end if
      end for
      if $t = 0$ **then**
        Update local static prototype $\tilde{Z}_l^t$ as Eq. (1) to (5) for all class $c \in \mathcal{C}_l^t$
      end if
    end for
    if $t = 0$ **then**
      Set unified prototype $Z_l^t = \tilde{Z}_G^t \cup \tilde{Z}_l^t$
    else
      Set unified prototype $Z_l^t = \tilde{Z}_G^t \cup \hat{Z}_l^t$
    end if
    Store local parameters $(P_l, \Phi_l,\ \Psi_l, Z_l^t)$
    Compute clients' weight ω_l^t
    Send local parameters $(P_l, \Phi_l,\ Z_l^t)$ and weight ω_l^t to server
    Server executes:
    if $R_T = 1$ **then**
      Receive clients initial static prototype $\tilde{Z}_l^t$ for $l \in [1..L]$
      Generate $Z_G^t = Z_G^t \cup Agg(\tilde{Z}_l^t$ for $l \in [1..L])$ and send Z_G^t to clients
    end if
    Receives selected clients S_l parameters $(P_l, \Phi_l,\ Z_l^t)$ and weight ω_l^t for $l \in [1..L]$
    Do weighted aggregation as in Eq. (16)
    Send global parameters (P_G, Φ_G, Z_G^t) to clients for the next round
  end for
end for
Output: Optimal Global parameters (P_G, Φ_G, Z_G)

Let $g(\Theta_l^k, \xi_l^k)$ be the unbiased estimator of ∇F_l, then the inequation above can be derived as:

$$\mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - F_l(\Theta_l^k) \le -\alpha_l^k \nabla ||F_l(\Theta_l^k)||_2^2 + {\alpha_l^k}^2 \frac{L}{2} \mathbb{E}_{\xi_l^k}[||g(\Theta_l^k, \xi_l^k)||_2^2] \tag{A4}$$

The inequation above guarantees SGD convergence as long as the stochastic directions and stepsize are chosen. We apply the restriction below to avoid the harm of the second term of the right side in the inequation above,

$$\mathbb{V}[g(\Theta_l^k, \xi_l^k)] = \mathbb{E}[||g(\Theta_l^k, \xi_l^k)||_2^2] - ||\mathbb{E}[g(\Theta_l^k, \xi_l^k)]||_2^2. \tag{A5}$$

Adopting first and second-moment limit as in [1], then we add the following assumption.

Assumption 6: The objective function F_l and SG satisfy the following conditions.

(a). The sequence of $\{\Theta_l^k\}$ is contained in an open space where F_l is bounded below by a scalar F_{inf}

(b) Exist scalars $\nu_G \ge \nu > 0$ so that for all $k \in \mathbb{N}$ satisfy:

$$\begin{aligned} &\nabla F_l(\Theta_l^k)^T \mathbb{E}_{\xi_l^k}[g(\Theta_l^k, \xi_l^k)] \ge \nu ||\nabla F_l(\Theta_l^k)T||_2^2, and \\ &||\mathbb{E}_{\xi_l^k}[g(\Theta_l^k, \xi_l^k)]||_2 \le \nu_G ||\nabla F_l(\Theta_l^k)||_2. \end{aligned} \tag{A6}$$

(c) Exist scalars $m_1 \ge 0$ and $m_2 \ge 0$ so that for all $k \in \mathbb{N}$ satisfy:

$$\mathbb{V}[g(\Theta_l^k, \xi_l^k)] \le m_1 + m_2 ||\nabla F_l(\Theta_l^k)||_2^2 \tag{A7}$$

Combining assumption 6 and restriction criteria as presented in equation (5), then we have:

$$\begin{aligned} \mathbb{E}_{\xi_l^k}[||g(\Theta_l^k, \xi_l^k)||_2^2] \le m_1 + m_G ||\nabla F_l(\Theta_l^k)||_2^2, with \\ m_G = m_2 + \nu_G^2 \ge \nu^2 > 0 \end{aligned} \tag{A8}$$

Then by substituting $\mathbb{E}_{\xi_l^k}[||g(\Theta_l^k, \xi_l^k)||_2^2]$ from equation (A8) into equation (A3), we have:

$$\begin{aligned} \mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - F_l(\Theta_l^k) \le &- \alpha_l^k \nabla F_l(\Theta_l^k)^T \mathbb{E}[g(\Theta_l^k, \xi_l^k)] \\ &+ {\alpha_l^k}^2 \frac{L}{2}(m_1 + m_G ||\nabla F_l(\Theta_l^k)||_2^2) \end{aligned} \tag{A9}$$

Assumption 5 ensures that $\{\alpha_l^k\} \to 0$ is practically achievable by applying a learning rate scheduler (with decay) that reduces the learning rate in each step of local training. Then by choosing $\alpha_l^k L m_G \le \nu$ and substituting $\nabla F_l(\Theta_l^k)^T \mathbb{E}[g(\Theta_l^k, \xi_l^k)]$ in equation (A9) with the condition in assumption 6.b, we have

$$\begin{aligned} \mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - F_l(\Theta_l^k) \le &- \alpha_l^k \nu ||\nabla F_l(\Theta_l^k)||_2^2 \\ &+ {\alpha_l^k}^2 \frac{L}{2}(m_1 + m_G ||\nabla F_l(\Theta_l^k)||_2^2) \end{aligned} \tag{A10}$$

Applying expectation into the equation above we get

$$\begin{aligned} \mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - \mathbb{E}[F_l(\Theta_l^k)] \le &- \alpha_l^k \nu \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] \\ &+ {\alpha_l^k}^2 \frac{1}{2}(m_1 + m_G \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2]) \\ \mathbb{E}_{\xi_l^k}[F_l(\Theta_l^{k+1})] - \mathbb{E}[F_l(\Theta_l^k)] \le &- \frac{1}{2}\nu \alpha_l^k \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] \\ &+ \frac{1}{2}{\alpha_l^k}^2 L m_1 \end{aligned} \tag{A11}$$

Sum both sides for $k \in \{1, ..., K\}$ we get

$$\begin{aligned} &F_{inf} - \mathbb{E}[F(\Theta_l^1)] \le \mathbb{E}[F_l(\Theta_l^{K+1})] - \mathbb{E}[F_l(\Theta_l^1)] \\ &F_{inf} - \mathbb{E}[F(\Theta_l^1)] \le -\frac{1}{2}\nu \sum_{k=1}^{K} \alpha_l^k \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] + \frac{1}{2} L m_1 \sum_{k=1}^{K} {\alpha_l^k}^2 \end{aligned} \tag{A12}$$

Dividing by ν for both sides, then we get

$$\sum_{k=1}^{K} \alpha_l^k \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] \le \frac{2(\mathbb{E}[F(\Theta_l^1)] - F_{inf})}{\nu} + \frac{L m_1}{\nu} \sum_{k=1}^{K} {\alpha_l^k}^2 \tag{A13}$$

Applying $\lim_{K\to\infty}$ and assumption 5 to the equation above we get

$$\begin{aligned} \lim_{K\to\infty} \sum_{k=1}^{K} \alpha_l^k \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] \le &\frac{2(\mathbb{E}[F(\Theta_l^1)] - F_{inf})}{\nu} \\ &+ \frac{L m_1}{\nu} \lim_{K\to\infty} \sum_{k=1}^{K} {\alpha_l^k}^2 < \infty \end{aligned} \tag{A14}$$

Dividing both sides with $\sum_{k=1}^{K} \alpha_l^k$, and following assumption 5 where $\lim_{K\to\infty} \sum_{k=1}^{K} \alpha_l^k = \infty$ and $\lim_{K\to\infty} \sum_{k=1}^{K} {\alpha_l^k}^2 < \infty$, then the right side will return 0. Therefore, we have

$$\lim_{K\to\infty} \frac{\sum_{k=1}^{K} \mathbb{E}[\alpha_l^k ||\nabla F_l(\Theta_l^k)||_2^2]}{\sum_{k=1}^{K} \alpha_l^k} = 0 \tag{A15}$$

$$\lim_{K\to\infty} \mathbb{E}[\frac{\sum_{k=1}^{K} \alpha_l^k ||\nabla F_l(\Theta_l^k)||_2^2}{\sum_{k=1}^{K} \alpha_l^k}] = 0 \tag{A16}$$

$$\lim_{k\to\infty} \mathbb{E}[||\nabla F_l(\Theta_l^k)||_2^2] = 0 \tag{A17}$$

The equation (A17) proves the convergence for local training in l-th client where the gradient of loss F converges to 0 along with the increase of training step/iteration k and the decreasing of learning rate α.

B. Proof of Theorem 2

Let the selected local clients $\{S_l\}_{l=1}^{l=L_S}$ are conduct local optimization with its local training data $\{\mathcal{T}_l^t \cup Z^t\}_{l=1}^{l=L_S}$ coordinated by central server S_G, where $\mathcal{T}_l^t$ is local training sample for client l for task t. Local training is conducted in k steps/iterations using a sample i.e. minibatch of local training set $\xi_l^k \in \mathcal{T}_l^t$ on each step. Global synchronization is executed in each round $r = \{1, 2, ..., R\}$. We global synchronization step as $\mathcal{I}_E = \{rE | r = 1, 2, ...R\}$. Following [2] we define Θ_l^{k+1} represents the local parameter of l-client after communication steps, while $\mathit{\Theta}_l^{k+1}$ represents the local parameter after an immediate result of one step SGD. Therefore the definition satisfies:

$$\mathit{\Theta}_l^{k+1} = \Theta_l^k - \alpha_l^k \nabla F_l(\Theta_l^k, \xi_l^k) \tag{A18}$$

$$\Theta_l^{k+1} = \begin{cases} \mathit{\Theta}_l^{k+1} \text{ if } k+1 \notin \mathcal{I}_E \\ \sum_{l=1}^{L_S} w_l^k \mathit{\Theta}_l^{k+1} \text{ if } k+1 \in \mathcal{I}_E \end{cases} \tag{A19}$$

Where $w_l = \omega_l / \sum_{l=1}^{L_S} \omega_l$, where ω_l is the weight of l-th client. We define $\bar{\mathit{\Theta}}_l^{k+1} = \sum_{l=1}^{L_S} w_l \mathit{\Theta}_l^{k+1}$ and $\bar{\Theta}_l^{k+1} =$

$\sum_{l=1}^{L_S} w_l \Theta_l^{k+1}$, $\bar{\mathcal{O}}_l^{k+1}$ is the result of single step SGD iteration from $\bar{\Theta}_l^{k+1}$. We also define $\bar{g}^k = \sum_{l=1}^{L_S} w_l \nabla F_l(\Theta_l^k)$ and $g^k = \sum_{l=1}^{L_S} w_l \nabla F_l(\Theta_l^k, \xi_l^k)$. We adopt the following lemmas from [2] where derived from fully participating clients in federated learning.

Lemma 1: By applying assumptions 1 and 2, in one step SGD training and chose $\alpha \leq \frac{1}{4L}$ we have $\mathbb{E}[||\bar{\mathcal{O}}^{k+1} - \Theta^*||^2] \leq (1-\alpha^k\mu)\mathbb{E}[||\bar{\Theta}^k - \Theta^*||^2] - (\alpha^k)^2\mathbb{E}[||g^k - \bar{g}^k||^2] + 6L(\alpha^k)^2\Gamma + 2\mathbb{E}[\sum_{l=1}^{L_S} w_l||\bar{\Theta}^k - \Theta_l^k||^2]$ where $\Gamma = F^* - \sum_{l=1}^{L_S} w_l F_l^* \geq 0$.

Lemma 2: By applying assumption 3, the gradient function follows:
$\mathbb{E}[||\bar{g}^k - \bar{g}^k||^2] \leq \sum_{l=1}^{L_S} w_l^2 \sigma_l^2$, where σ_l^2 is the variance of Θ_l

Lemma 3: By applying assumption 4, where α^k is non-increasing and it satisfies $\alpha^k \leq \alpha^{k+E}$for all$k \geq 0$, then we have $\mathbb{E}[\sum_{l=1}^{L_S} ||\bar{\Theta}^{k+1} - \Theta_l^k||^2] \leq 4(\alpha^k)^2(E-1)^2G^2$

In fully participating clients we always have $\bar{\Theta}^{k+1} = \bar{\mathcal{O}}^{k+1}$. However, in partially participating clients we use a random sampling mechanism so that It satisfies $\mathbb{E}_{S_L}[\bar{\Theta}^{k+1}] = \bar{\mathcal{O}}^{k+1}$. We also adopt the bounding condition from [2] as shown in lemma 4.

Lemma 4: The expected different between $\bar{\Theta}^{k+1}$and $\bar{\mathcal{O}}^{k+1}$ bounded by : $\mathbb{E}_{S_L}[||\bar{\mathcal{O}}^{k+1} - \bar{\Theta}^{k+1}||^2] \leq \frac{4}{L_S}(\alpha^k)^2E^2G^2$ and in the case of w_l is uniform for all l-th client, then $\mathbb{E}_{S_L}[||\bar{\mathcal{O}}^{k+1} - \bar{\Theta}^{k+1}||^2] \leq \frac{4(N_S - L_S)}{N_S - 1}(\alpha^k)^2E^2G^2$, where N_S is total clients and L_S is number of selected clients.

Please note that

$$||\bar{\Theta}^{k+1} - \Theta^*||^2 = ||\bar{\Theta}^{k+1} - \Theta^*||^2 \tag{A20}$$

$$||\bar{\Theta}^{k+1} - \Theta^*||^2 = ||\bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1} + -\bar{\mathcal{O}}^{k+1} - \Theta^*||^2 \tag{A21}$$

$$\begin{aligned} ||\bar{\Theta}^{k+1} - \Theta^*||^2 =& ||\bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1}||^2 + ||\bar{\mathcal{O}}^{k+1} - \Theta^*||^2 \\ &+ 2||\bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1}||.||\bar{\mathcal{O}}^{k+1} - \Theta^*|| \end{aligned} \tag{A22}$$

$$\begin{aligned} ||\bar{\Theta}^{k+1} - \Theta^*||^2 =& ||\bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1}||^2 + ||\bar{\mathcal{O}}^{k+1} - \Theta^*||^2 \\ &+ 2\langle \bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1}, \bar{\mathcal{O}}^{k+1} - \Theta^* \rangle \end{aligned} \tag{A23}$$

In the case of $k+1 \notin \mathcal{I}_E$, then the term $||\bar{\Theta}^{k+1} - \bar{\mathcal{O}}^{k+1}||^2$ vanishes. Then by applying lemma 4, we get

$$\mathbb{E}[||\bar{\Theta}^{k+1} - \Theta^*||^2] \leq (1-\alpha^k\mu)\mathbb{E}[||\bar{\Theta}^{k+1} - \Theta^*||^2] + (\alpha^k)B \tag{A24}$$

In the case of $k+1 \in \mathcal{I}_E$, then by applying lemma 4, we get

$$\mathbb{E}[||\bar{\Theta}^{k+1} - \Theta^*||^2] \leq (1-\alpha^k\mu)\mathbb{E}[||\bar{\Theta}^{k+1} - \Theta^*||^2] + (\alpha^k)(B+C) \tag{A25}$$

where $B = \sum_{l=1}^{L_S} w_l\sigma_l^2 + 6L\Sigma + 8(E-1)^2G^2$ and $C = \frac{4(N_S - L_S)}{N_S - 1}(E^2G^2)$ if w_l is uniform and $C = \frac{4}{L_S}(E^2G^2)$ otherwise.

By choosing $\alpha^k = \frac{\beta}{k+\delta}$ for some $\beta > 1/\mu$ and $\delta > 0$ so that $\alpha^1 \leq \min\{1/\mu, 1/4L\} = 1/4L$ and $\alpha^k \leq 2\alpha^{k+E}$ then we have $\mathbb{E}[||\bar{\Theta}^{k+1} - \Theta^*||^2] \leq \frac{v}{\delta+k}$ where $v = \max\{\frac{\beta^2(B+C)}{\beta\mu - 1}, (\delta + 1)||\bar{\Theta}^{k+1} - \Theta^*||^2\}$

Then, by applying a strong convexity assumption of F we have

$$\mathbb{E}[\bar{\Theta}^k] - F^* \leq \frac{L}{2}\Delta^k \leq \frac{L}{2}\frac{v}{\delta+k} \tag{A26}$$

where F^* is the minimum value of F where optimum parameter Θ^* is achieved. Later on, if we choose $\beta = 2/\mu, \delta = \max\{8L/\mu, E\}$and denote $\kappa = L/\mu, \alpha^k = 2/u(1/(\delta+k))$ then we have

$$\mathbb{E}[F(\bar{\Theta}^k)] - F^* \leq \frac{\kappa}{(\delta+k-1)}(\frac{2(B+C)}{\mu} + \frac{\mu\delta}{2}\mathbb{E}||\Theta^1 - \Theta^*||) \tag{A27}$$

The equation generalizes federated learning where the model is trained in a total of k steps/iterations where practical implementation satisfies $k = b.E.R, b$ is the number of batches. here we know that $k > R$ as E and b are positive integers. Therefore, substituting k with R in the inequation above will produce a higher amount of the right side. Therefore, the inequation above can be generalized into:

$$\mathbb{E}[F(\Theta^R)] - F^* \leq \frac{\kappa}{(\delta+R-1)}(\frac{2(B+C)}{\mu} + \frac{\mu\delta}{2}\mathbb{E}||\Theta^1 - \Theta^*||) \tag{A28}$$

The equation above can be derived into:

$$\mathbb{E}[F(\Theta^R)] - F^* \leq \frac{1}{R}\frac{\kappa}{(\delta/R+1-1/R)}(\frac{2(B+C)}{\mu} + \frac{\mu\delta}{2}\mathbb{E}||\Theta^1 - \Theta^*||) \tag{A29}$$

Let $A = \frac{\kappa}{(\delta/R+1-1/R)}$ is a positive number. Then the equation above can be derived into:

$$\mathbb{E}[F(\Theta^R)] - F^* \leq \frac{A}{R}(\frac{2(B+C)}{\mu} + \frac{\mu\delta}{2}\mathbb{E}||\Theta^1 - \Theta^*||) \tag{A30}$$

The inequation (A30) guarantees the proposed weighted federated learning achieves a convergence condition that is upper bounded by the amount on the right side.

C. Proof of Theorem 3

Given Θ^* and Θ are optimal parameter in $\mathcal{T}_l^t \cup Z^t$ and $\mathcal{T}_l^t$ respectively, where $\mathcal{T}_l^t \subset \mathcal{T}^t$, where $|\mathcal{T}_l^t|/|\mathcal{T}^t| = \eta \in (0,1)$, then we have

$$F(\Theta; \mathcal{T}^t) = \eta F(\Theta; \mathcal{T}_l^t) + (1-\eta)F(\Theta; (\mathcal{T}^t - \mathcal{T}_l^t)) \tag{A31}$$

$$F(\Theta^*; \mathcal{T}^t) = \eta F(\Theta^*; \mathcal{T}_l^t) + (1-\eta)F(\Theta^*; (\mathcal{T}^t - \mathcal{T}_l^t)) \tag{A32}$$

Suppose that Θ^o is the initial value both for $\Theta$and Θ^* that set by random uniform initiation method. Therefore for all class $c \in \mathcal{T}_c^t = \mathcal{T}_{y=c}^t$ It satisfy $F(\Theta^o; \mathcal{T}_c^t) = e^o$. After optimally learning on $\mathcal{T}_l^t$ and $\mathcal{T}_l^t \cup Z^t$ then Θ^o become to Θ and Θ^* respectively. Please note that Θ learns only available class in $\mathcal{T}_l^t$, while Θ^* learns classes that available in $\mathcal{T}_l^t$ and classes in $\mathcal{T}^t - \mathcal{T}_l^t$ via Z^t. Suppose that the loss for predicting classes in $\mathcal{T}_l^t$ is defined as $e^a < e^o$ then we have $F(\Theta; \mathcal{T}_l^t) = F(\Theta^*; \mathcal{T}_l^t) = e^a < e^o$. Since the backbone is frozen, Θ^* learn Z^t then we have $F(\Theta; (\mathcal{T}^t - \mathcal{T}_l^t)) = e^o$, while $F(\Theta^*; (\mathcal{T}^t - \mathcal{T}_l^t)) = e^b$, where $e^a \geq e^b \geq e^o$.

Then equation (A31) and (A32) can be derived to

$$F(\Theta; \mathcal{T}^t) = \eta e^a + (1-\eta)e^o \tag{A33}$$

$$F(\Theta^*; \mathcal{T}^t) = \eta e^a + (1-\eta)e^b \tag{A34}$$

Substracting the equations above, then we have

$$F(\Theta; \mathcal{T}^t) - F(\Theta^*; \mathcal{T}^t) = \eta e^a + (1-\eta)e^o - (\eta e^a + (1-\eta)e^b) \tag{A35}$$

$$F(\Theta;\mathcal{T}^t)-F(\Theta^*;\mathcal{T}^t)=\eta e^a+(1-\eta)e^o-\eta e^a-(1-\eta)e^b \quad (A36)$$

$$F(\Theta;\mathcal{T}^t)-F(\Theta^*;\mathcal{T}^t)=(1-\eta)e^o-(1-\eta)e^b \quad (A37)$$

$$F(\Theta;\mathcal{T}^t)-F(\Theta^*;\mathcal{T}^t)=(1-\eta)(e^o-e^b) \quad (A38)$$

Since $0<\eta<1$ and $e^o>e^b$ the right side of the inequation above has a positive value. By choosing a small positive value $\epsilon>0$ where $(1-\eta)(e^o-e^b)\geq\epsilon$ then we have.

$$F(\Theta;\mathcal{T}^t)-F(\Theta^*;\mathcal{T}^t)\geq\epsilon \quad (A39)$$

Inequation above proves that Θ^* is more generalized to $\mathcal{T}^t$ than Θ. This shows that our idea i.e. empowering prompt learning with shared unified prototypes improves model generalization.

III. Detailed Experiment Setting

Experimental Details: Our numerical study is executed under a single NVIDIA A100 GPU with 40 GB memory across 3 runs with different random seeds {2023,2024,2025}. Fed-L2P, Fed-DualP, Fed-CODAP, and UOPP train T number of prompts $P\in R^{5\times 768}$ and $\Phi\in R^{|\mathcal{C}\times 768|}$ head layer, while the competitors train whole CNN models following their original implementation. Adapting from [3] with our computational resources, each experiment is simulated by 20 total clients and 1 global server, where in each round, 6 (30%) local clients are selected randomly. Each client randomly receives 60% ($\eta=0.6$) class label space. The total global round is set to 90 (10 rounds per class) for CIFAR100 and MiniImageNet and 110 for CUB200. The local training on each client is set with maximum of 20 epochs, and the learning rate is set by choosing the best value from $\{0.02, 0.002\}$ for the base task, and $\{0.1, 0.2, 0.3\}$ for the few shot task.

Performance Metric: On each task, we evaluate the consolidated algorithms with (Acc(.)) accuracy metrics. Besides, we also measure the accuracy of base classes, novel classes and harmonic mean accuracy. Base classes are the classes that belong to the base task i.e. $c\in\mathcal{C}^0$, while novel classes are the classes that belong to task 1 until the current task i.e. $c\in\{\mathcal{C}^1\cup...\cup\mathcal{C}^t\}$. Harmonic mean accuracy is defined as $(Acc(BaseClasses)+Acc(NovelClasses))/2$. The harmonic mean indicates the balance between the performance base classes and novel classes, in other words, it represents stability-plasticity performance. We also measure performance drop (PD), that is the accuracy difference between the first task, and the last task.

IV. Detailed Numerical Results on Benchmark Datasets

In this section, we present the detailed numerical result as shown in Tables A1, A2, and A3.

V. Detailed Numerical Results on Stability-Plasticity Analysis

In this section, we present the detailed numerical results on the stability-plasticity analysis of UOPP as shown in Tables A4, A5, and A6.

VI. Detailed Numerical Results Different Local Clients and Global Rounds

In this section we present the detailed numerical results on different local clients and rounds as presented in tables A7 and A8

VII. Detailed Numerical Results of Ablation Study

In this section we present detailed numerical results on the ablation study as shown in table A9.

VIII. Detailed Compelxity Analysis

Following the pseudo-code in Algorithm 1, UOPP have several operations e.g. generate static prototype (line 11, 24, 34), drawing $\mathcal{S}$, $\mathcal{Q}$ from prototypes (line 25), Rectify prototype (line 26), updating model parameters (line 20-22, 28-30), forming unified prototype (line 27, 38, 40) data exchange between clients and server. Knowing that accumulating on all batches, generating prototype or compute features from $\mathcal{T}_l^t$ cost $O(N_l^t)$, drawing (augment) samples from feature costs $O(N_l^t)$, rectifying prototypes cost costs $O(N_l^t)$, parameters update cost costs $O(N_l^t)$, forming uniform prototype cost $O(1)$, and parameters exchange include aggregation costs $O(1)$, and we have 1 base task and T few-shot tasks (total task is (T+1)) then the UOPP complexity will be:

$$O(UPPP)=O(BaseTask)+O(FewStotTask) \quad (A40)$$

$$\begin{aligned}O(UOPP)=&O(1)+R_T(O(clients_{base})+O(server_{base})\\&+O(1)+T.R_T.(O(clients_{fs})+O(server_{fs}))\end{aligned} \quad (A41)$$

$$\begin{aligned}O(UOPP)=&O(1)+R_T.(L.O(1client_{base})+O(server_{base}))\\&+T.R_T.(L.O(1client_{fs})+O(server_{fs}))\end{aligned} \quad (A42)$$

$$\begin{aligned}O(UOPP)=&O(1)+R_T.(L(O(N_l^0)+O(E.N_l^0)\\&+O(E.N_l^0))+O(1)\\&+T.R_T.(L(O(N_l^t)+O(E.N_l^t)+O(E.N_l^t)\\&+O(E.1)+O(E.N_l^t))+O(1)\end{aligned} \quad (A43)$$

$$O(UOPP)=O(1)+R_T.L.O(E.N_l^0)+T.R_T.L.O(E.N_l^t) \quad (A44)$$

$$O(UOPP)=O(1)+R_T.L.E.O(N_l^0)+T.R_T.L.E.O(N_l^t) \quad (A45)$$

$$O(UOPP)=O(1)+R_T.L.E(O(N_l^0)+T.O(N_l^t)) \quad (A46)$$

Please note that $N_l=N_l^0+N_l^1+...+N_l^T=N_l^0+T(N_l^t), t\in[1..T]$. Therefore, the equation above can be derived into:

$$O(UOPP)=O(1)+R_T.L.E(O(N_l^0+T.ON_l^t)) \quad (A47)$$

$$O(UOPP)=R_T.L.E.O(N_l) \quad (A48)$$

$$O(UOPP)=O(R_T.L.E.N_l) \quad (A49)$$

Method	Accuracy in each session (%)											Avg	PD	Gap
	0	1	2	3	4	5	6	7	8	9	10			
Fed-L2P	73.7	67.8	62.1	58.9	61.1	57.0	52.7	50.3	48.4	47.0	49.4	57.1	24.3	19.6
Fed-DualP	78.2	76.4	71.2	66.3	65.6	62.7	58.4	54.3	51.7	50.5	51.3	62.4	26.9	14.3
Fed-CODAP	73.1	56.5	48.8	39.6	37.5	35.2	32.6	30.8	28.3	28.1	27.4	39.8	45.7	36.9
Fed-S3C	18.6	18.2	17.3	15.6	14.9	13.6	13.2	12.6	11.8	11.6	10.8	14.4	7.9	62.3
TARGET	29.0	27.0	24.8	22.9	21.2	19.8	18.6	17.5	16.5	15.6	14.9	20.7	14.2	56.0
LGA	17.1	9.3	7.1	6.9	7.1	6.8	6.6	6.5	6.3	5.9	5.7	7.8	11.4	69.0
UOPP	85.9	84.7	83.2	80.2	79.2	76.6	74.0	72.7	71.1	70.3	66.3	76.7	19.5	0.0

TABLE A1

NUMERICAL RESULT OF THE CONSOLIDATED ALGORITHMS IN CUB200 DATASET WITH 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Method	S	Accuracy in each session (%)									Avg	PD	Gap
		0	1	2	3	4	5	6	7	8			
Fed-L2P	5	73.47	74.20	73.37	71.88	70.85	70.72	69.28	68.66	68.37	71.20	5.10	18.81
Fed-DualP	5	76.39	82.75	83.37	80.80	79.93	78.26	77.73	76.98	77.11	79.26	-0.72	10.75
Fed-CODAP	5	81.73	69.29	70.81	68.67	67.17	66.14	64.32	64.79	64.12	68.56	17.62	21.45
Fed-S3C	5	44.51	48.97	47.77	45.35	43.48	41.47	40.33	39.32	37.71	43.21	6.80	46.80
TARGET	5	68.90	63.61	59.06	55.12	51.68	48.64	45.94	43.52	41.34	53.09	27.56	36.92
LGA	5	73.76	69.80	65.59	60.26	56.87	52.94	50.66	47.69	44.89	58.05	28.87	31.96
UOPP	5	90.57	90.58	90.85	90.96	91.23	91.51	91.56	91.74	81.05	90.01	9.52	0.00
Fed-L2P	5	77.00	74.91	74.73	74.40	73.45	74.20	73.22	73.46	73.25	74.29	3.75	14.33
Fed-DualP	1	78.66	85.01	85.31	83.60	82.58	81.59	81.04	80.24	79.53	81.95	-0.87	6.67
Fed-CODAP	1	83.29	73.27	72.54	68.37	67.52	65.54	62.11	64.19	61.67	68.72	21.62	19.90
Fed-S3C	1	44.51	48.70	46.76	44.26	42.09	40.11	38.51	37.35	35.44	41.97	9.07	46.65
TARGET	1	68.90	63.61	59.06	55.12	51.68	48.64	45.94	43.52	41.34	53.09	27.56	35.53
LGA	1	73.58	67.00	63.44	58.88	56.90	54.00	52.82	49.25	48.78	58.29	24.80	30.33
UOPP	1	90.65	90.16	89.97	89.49	89.53	88.29	87.66	87.04	84.82	88.62	5.83	0.00

TABLE A2

NUMERICAL RESULT OF THE CONSOLIDATED ALGORITHMS IN CIFAR100 DATASET WITH 5-SHOT AND 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Method	S	Accuracy in each session (%)									Avg	PD	Gap
		0	1	2	3	4	5	6	7	8			
Fed-L2P	5	81.02	78.22	78.66	79.44	79.67	78.14	77.65	77.48	80.00	78.92	1.0	14.0
Fed-DualP	5	83.93	89.31	88.11	87.47	87.23	84.97	83.96	83.78	84.38	85.91	-0.4	7.0
Fed-CODAP	5	90.21	83.35	82.31	80.27	78.89	77.79	77.13	75.94	75.09	80.11	15.1	12.8
Fed-S3C	5	31.91	32.97	31.74	30.96	29.93	28.92	27.66	27.06	26.43	29.73	5.5	63.2
TARGET	5	58.10	53.64	49.80	46.48	43.58	41.02	38.74	36.70	34.86	44.77	23.2	48.2
LGA	5	46.28	29.54	13.94	12.58	11.34	10.67	9.72	9.14	8.67	16.88	37.6	76.0
UOPP	5	93.65	93.24	92.97	92.60	92.73	92.92	92.49	92.73	92.92	92.92	0.7	0.0
Fed-L2P	5	83.02	79.99	79.92	79.54	80.20	80.84	80.55	80.60	82.57	80.80	0.4	12.1
Fed-DualP	1	85.47	90.06	88.77	88.44	88.14	86.22	85.04	84.98	84.80	86.88	0.7	6.0
Fed-CODAP	1	90.94	83.65	81.53	80.21	79.50	77.48	76.71	75.89	75.24	80.13	15.7	12.8
Fed-S3C	1	32.84	33.19	31.83	30.69	29.17	27.90	26.54	25.68	24.60	29.16	8.2	63.8
TARGET	1	58.10	53.64	49.80	46.48	43.58	41.02	38.74	36.70	34.86	44.77	23.2	48.2
LGA	1	46.14	15.72	14.45	13.23	12.36	11.43	8.16	9.35	7.25	15.34	38.9	77.6
UOPP	1	93.66	93.15	92.72	92.04	92.20	92.05	91.03	91.56	91.48	92.21	2.2	0.7

TABLE A3

NUMERICAL RESULT OF THE CONSOLIDATED ALGORITHMS IN MINIIMAGENET DATASET WITH 5-SHOT AND 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Since E is set as a small constant in our method i.e. 1-20 and $R_T < R$, then the UOPP complexity will be:

$$O(UOPP) = O(R.L.N_l) \tag{A50}$$

Our derivation shows that the baseline and our proposed method (PIP) have the same complexity i.e. $O(R.L.N_l)$ where R is total global rounds, L is the number of selected local clients in each round and N_l is the number of samples in each client.

REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[2] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[3] J. Dong, H. Li, Y. Cong, G. Sun, Y. Zhang, and L. V. Gool, "No one left behind: Real-world federated class-incremental learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2023.

Method	S	Base Classes Accuracy in each session (%) 0	1	2	3	4	5	6	7	8	9	10	Avg	PD	Gap
Fed-L2P	5	73.2	71.0	70.6	72.6	75.2	74.0	74.4	74.9	74.3	73.5	75.0	73.5	-1.8	11.7
Fed-DualP	5	79.0	78.2	79.2	79.8	79.6	79.3	79.3	79.4	78.9	79.4	80.1	79.3	-1.1	5.9
Fed-CODAP	5	71.7	49.7	37.5	28.2	29.2	24.3	23.9	22.7	21.9	22.0	21.5	32.1	50.2	53.1
Fed-S3C	5	18.6	18.7	19.7	18.8	18.2	17.8	18.2	18.0	18.3	17.8	17.4	18.3	1.3	66.9
TARGET	5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	0.0	54.7
LGA	5	16.8	7.6	7.5	6.4	7.1	7.3	7.7	7.4	6.9	6.8	7.2	8.1	9.6	77.1
UOPP	5	86.2	85.5	85.3	85.3	84.9	84.9	84.9	84.9	84.9	85.1	85.2	85.2	1.0	0.0
Fed-L2P	1	73.7	68.9	69.0	70.8	74.1	73.3	72.8	73.3	73.0	72.5	76.1	72.5	-2.4	12.3
Fed-DualP	1	78.2	75.6	76.8	77.7	77.4	77.1	77.0	76.0	76.4	75.9	77.6	76.9	0.6	7.9
Fed-CODAP	1	73.1	53.8	46.0	38.9	36.5	33.9	32.1	30.2	28.7	28.5	26.8	38.9	46.3	45.8
Fed-S3C	1	18.6	18.8	19.4	18.7	18.3	17.5	18.1	17.7	17.8	17.9	17.4	18.2	1.2	66.6
TARGET	1	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7	0.0	55.1
LGA	1	17.1	6.6	3.7	4.1	4.9	3.8	3.7	3.9	3.6	3.3	3.7	5.3	13.3	79.5
UOPP	1	85.9	85.3	85.2	85.2	85.2	85.2	84.6	84.6	84.6	84.4	82.3	84.8	3.6	0.0

TABLE A4

BASE CLASSES ACCURACY OF THE CONSOLIDATED ALGORITHMS IN CUB200 DATASET WITH 5-SHOT AND 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Method	S	Novel Classes Accuracy in each session (%) 1	2	3	4	5	6	7	8	9	10	Avg	PD	Gap
Fed-L2P	5	54.24	26.38	20.33	26.23	22.57	18.65	19.95	23.51	23.21	26.67	26.2	27.6	47.0
Fed-DualP	5	73.84	37.63	25.81	31.64	29.34	23.79	21.30	19.07	19.92	22.12	30.4	51.7	42.8
Fed-CODAP	5	86.98	66.49	48.03	47.08	40.15	41.90	40.50	38.19	38.16	37.18	48.5	49.8	24.7
Fed-S3C	5	16.73	9.07	5.94	8.48	7.23	6.42	6.45	5.89	6.27	5.64	7.8	11.1	65.4
TARGET	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.0	73.2
LGA	5	43.24	25.67	19.30	15.51	11.81	10.47	9.83	8.63	7.48	7.60	16.0	35.6	57.2
UOPP	5	90.32	83.10	75.39	73.48	68.88	68.15	68.67	68.87	69.96	65.61	73.2	24.7	0.0
Fed-L2P	1	56.63	26.80	19.52	29.01	25.22	19.75	18.04	18.39	19.11	23.28	25.6	33.4	35.6
Fed-DualP	1	84.71	42.93	28.67	36.63	34.52	27.94	23.79	21.48	22.78	25.56	34.9	59.1	26.3
Fed-CODAP	1	84.95	63.19	42.17	39.81	37.57	33.37	31.69	27.88	27.67	27.95	41.6	57.0	19.6
Fed-S3C	1	12.54	6.77	5.56	6.53	6.06	5.26	5.60	4.51	4.70	4.32	6.2	8.2	55.0
TARGET	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	0.0	61.2
LGA	1	36.49	24.47	16.66	12.90	12.86	11.40	10.26	9.71	8.73	7.72	15.1	28.8	46.1
UOPP	1	78.61	72.91	63.58	64.38	59.64	56.67	56.03	54.47	54.84	50.76	61.2	27.9	0.0

TABLE A5

NOVEL CLASSES ACCURACY OF THE CONSOLIDATED ALGORITHMS IN CUB200 DATASET WITH 5-SHOT AND 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Method	S	Harmonic Mean Accuracy in each session (%) 1	2	3	4	5	6	7	8	9	10	Avg	PD	Gap
Fed-L2P	5	62.61	48.50	46.45	50.71	48.31	46.52	47.41	48.92	48.35	50.84	49.9	11.8	29.3
Fed-DualP	5	76.03	58.42	52.81	55.63	54.34	51.55	50.34	48.99	49.68	51.10	54.9	24.9	24.3
Fed-CODAP	5	68.35	51.98	38.12	38.15	32.25	32.91	31.59	30.05	30.07	29.35	38.3	39.0	40.9
Fed-S3C	5	17.72	14.40	12.39	13.36	12.54	12.29	12.23	12.08	12.03	11.50	13.1	6.2	66.1
TARGET	5	15.26	15.26	15.26	15.26	15.26	15.26	15.26	15.26	15.26	15.26	15.3	0.0	63.9
LGA	5	25.43	16.61	12.87	11.29	9.57	9.07	8.61	7.75	7.16	7.39	11.6	18.0	67.6
UOPP	5	87.92	84.21	80.35	79.21	76.89	76.54	76.80	76.89	77.54	75.40	79.2	12.5	0.0
Fed-L2P	1	62.75	47.91	45.16	51.56	49.25	46.29	45.66	45.68	45.82	49.70	49.0	13.0	23.9
Fed-DualP	1	80.16	59.88	53.18	57.02	55.80	52.45	49.91	48.92	49.35	51.56	55.8	28.6	17.1
Fed-CODAP	1	69.36	54.58	40.51	38.16	35.74	32.72	30.93	28.28	28.08	27.38	38.6	42.0	34.3
Fed-S3C	1	15.66	13.10	12.12	12.39	11.79	11.67	11.63	11.14	11.28	10.87	12.2	4.8	60.7
TARGET	1	14.86	14.86	14.86	14.86	14.86	14.86	14.86	14.86	14.86	14.86	14.9	0.0	58.0
LGA	1	21.53	14.07	10.37	8.89	8.33	7.55	7.05	6.64	6.03	5.73	9.6	15.8	63.3
UOPP	1	81.93	79.05	74.39	74.79	72.43	70.65	70.34	69.54	69.63	66.52	72.9	15.4	0.0

TABLE A6

HARMONIC MEAN ACCURACY OF THE CONSOLIDATED ALGORITHMS IN CUB200 DATASET WITH 5-SHOT AND 1-SHOT SETTING ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE GAP BETWEEN THE RESPECTED METHOD TO OUR PROPOSED METHOD (UOPP).

Method	L	Accuracy in each session (%)									Avg	PD
		0	1	2	3	4	5	6	7	8		
Fed-DualP	4	64.90	75.14	80.11	78.63	79.35	77.79	76.84	76.67	76.05	76.17	-11.15
Fed-DualP	6	77.02	82.98	83.90	81.00	80.18	78.85	78.33	77.55	77.75	79.73	-0.73
Fed-DualP	8	84.65	84.77	83.90	80.41	78.56	76.64	75.93	74.60	73.58	79.23	11.07
S3C	4	42.63	50.02	48.77	46.47	44.56	42.72	41.53	40.39	38.39	43.94	4.24
S3C	6	43.57	48.75	47.26	44.68	42.75	40.98	39.62	38.68	37.03	42.59	6.54
S3C	8	43.43	48.60	46.90	44.81	43.16	41.34	40.33	38.84	37.21	42.74	6.22
TARGET	4	66.75	61.62	57.21	53.40	50.06	47.12	44.50	42.16	40.05	51.43	26.70
TARGET	6	69.38	64.05	59.47	55.51	52.04	48.98	46.26	43.82	41.63	53.46	27.75
TARGET	8	73.53	67.88	63.03	58.83	55.15	51.91	49.02	46.44	44.12	56.66	29.41
LGA	4	72.98	68.8	62.81	57.57	54.29	51.51	49.81	46.67	42.58	56.34	30.40
LGA	6	73.6	71.4	65.86	59.45	56.71	52.92	50.23	47.88	44.34	58.04	29.26
LGA	8	73.73	70.03	65.9	60.4	56.31	52.68	50.37	47.32	44.61	57.93	29.12
UOPP	4	89.18	89.49	89.57	89.91	90.35	89.99	89.23	88.88	84.96	89.06	4.22
UOPP	6	90.10	90.34	90.63	90.80	91.21	91.62	91.63	91.75	79.20	89.70	10.90
UOPP	8	90.93	90.91	91.40	91.61	91.74	91.78	91.26	91.36	90.90	91.32	0.03

TABLE A7

ACCURACY OF THE CONSOLIDATED ALGORITHMS IN CIFAR100 DATASET WITH 5-SHOT SETTING ON DIFFERENT NUMBER OF SELECTED LOCAL CLIENTS ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND L INDICATES THE NUMBER OF SELECTED LOCAL CLIENTS.

Method	R	Accuracy in each session (%)									Avg	PD
		0	1	2	3	4	5	6	7	8		
Fed-DualP	54	79.40	81.40	83.44	82.67	82.39	81.85	81.31	80.78	80.32	81.51	-0.92
Fed-DualP	72	78.85	83.25	83.96	81.61	80.28	79.73	78.63	78.15	77.06	80.17	1.79
Fed-DualP	90	77.02	82.98	83.90	81.00	80.18	78.85	78.33	77.55	77.75	79.73	-0.73
S3C	54	43.42	49.51	48.01	45.41	43.69	41.96	40.89	39.64	38.10	43.40	5.32
S3C	72	51.20	53.95	51.97	49.09	47.10	45.11	43.74	42.88	41.13	47.35	10.07
S3C	90	43.57	48.75	47.26	44.68	42.75	40.98	39.62	38.68	37.03	42.59	6.54
TARGET	54	57.60	53.17	49.37	46.08	43.20	40.66	38.40	36.38	34.56	44.38	23.04
TARGET	72	67.28	62.11	57.67	53.83	50.46	47.49	44.86	42.49	40.37	51.84	26.91
TARGET	90	69.38	64.05	59.47	55.51	52.04	48.98	46.26	43.82	41.63	53.46	27.75
LGA	54	69.57	62.32	60.94	61.41	57.44	53.71	49.76	51.24	47.51	57.10	22.06
LGA	72	68.35	66.26	61.61	58.15	54.75	51.55	49.21	45.19	42.08	55.24	26.27
LGA	90	73.60	71.40	65.86	59.45	56.71	52.92	50.23	47.88	44.34	58.04	29.26
UOPP	54	89.87	89.75	90.33	90.73	90.91	91.12	91.33	91.18	91.32	90.73	-1.45
UOPP	72	90.37	90.29	90.70	90.99	90.94	91.34	91.27	91.07	90.13	90.79	0.24
UOPP	90	90.10	90.34	90.63	90.80	91.21	91.62	91.63	91.75	79.20	89.70	10.90

TABLE A8

ACCURACY OF THE CONSOLIDATED ALGORITHMS IN CIFAR100 DATASET WITH 5-SHOT SETTING ON DIFFERENT NUMBER OF ROUNDS ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND R INDICATES THE NUMBER OF ROUNDS.

Conf.	Harmonic Accuracy in each session (%)									Avg	PD	Gap
	0	1	2	3	4	5	6	7	8			
A (w/o St.Proto)	84.37	82.54	80.91	80.56	80.29	80.54	80.70	79.03	76.61	80.62	7.76	9.39
B (w/o Dy.Proto)	90.27	87.38	85.67	84.40	84.69	84.84	85.24	85.16	80.21	85.32	10.06	4.69
C (w/o FC Cls.)	88.25	88.66	89.07	89.16	89.63	90.11	90.27	90.62	82.76	88.72	5.49	1.29
D (w/o PB Cls.)	90.10	83.17	77.23	72.08	67.58	63.60	60.07	56.91	52.34	69.23	37.76	20.78
UOPP	90.57	90.58	90.85	90.96	91.23	91.51	91.56	91.74	81.05	90.01	9.52	0.00

TABLE A9

ACCURACY OF DIFFERENT CONFIGURATIONS IN CIFAR100 DATASET WITH 5-SHOT SETTING ON ACROSS 3 DIFFERENT SEEDED RUNS. S INDICATES THE NUMBER OF SHOTS FOR THE FEW SHOT TASKS, PD INDICATES THE PERFORMANCE DROP, AND GAP INDICATES THE DIFFERENCE ACCURACY TO PIP.