FEDPAC: CONSISTENT REPRESENTATION LEARNING FOR FEDERATED UNSUPERVISED LEARNING UNDER DATA HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated unsupervised learning enables collaborative model training on decentralized unlabeled data but faces critical challenges under data heterogeneity, which often leads to representation collapse from weak supervisory signals and semantic misalignment across clients. Without a consistent semantic structure constraints, local models learn disparate feature spaces, and conventional parameter averaging fails to produce a coherent global model. To address these issues, we propose Federated unsupervised learning with Prototype Anchored Consensus (FedPAC), a novel framework that establishes a consistent representation space via a set of learnable prototypes. FedPAC introduces a dual-alignment objective during local training: a semantic alignment loss that steers local models towards a prototype-anchored consensus to ensure cross-client semantic consistency, coupled with a representation alignment loss that promotes the learning of discriminative and stable features. On the server, prototypes are aggregated by an optimization-based strategy that preserves semantic knowledge and ensure the prototypes remain representative. We provide a rigorous convergence analysis for our method, formally proving its convergence under mild assumptions. Extensive experiments on benchmarks including CIFAR-10 and CIFAR-100 demonstrate that FedPAC significantly outperforms state-of-the-art methods across a wide range of heterogeneous settings.

1 INTRODUCTION

Federated learning (FL) (McMahan et al., 2017) enables a set of distributed clients to collaboratively train a shared model without exchanging raw data, thereby providing privacy preservation. A central challenge in FL is data heterogeneity: clients typically hold non-IID local datasets, and such distributional skew can make local updates to conflict, degrade the aggregated global model, and destabilize convergence. Existing FL algorithms generally assume supervised local training with abundant, high-quality labels. However, it is often impractical to collect large-scale, accurately annotated datasets in many practical applications. This pervasive label scarcity not only limits attainable performance but also undermines generalization to new domains, motivating methods that exploit the large volumes of unlabeled data distributed across clients. In this work, we study representation learning for federated unsupervised learning with non-IID data, aiming to extract robust and generalizable representations from distributed, unlabeled, and imbalanced data.

Federated unsupervised learning currently faces two fundamental challenges that impede the training of a high-quality global model. The first is **representation collapse**, where the weak supervisory signals from unlabeled data can lead to degenerate features with limited discriminability. Data heterogeneity further exacerbates training instability, increasing the risk of representation collapse. The second, and more complex challenge is **cross-client semantic misalignment**. Data heterogeneity across clients undermines the objective of learning a unified global representation, as each client learns distinct feature spaces tailored to its local data distribution. This causes representations of semantically similar samples to drift to disparate regions of the global feature space. This misalignment is often exacerbated by simple parameter averaging, which can blur semantic boundaries and paradoxically degrade the performance of global model. These issues expose a fundamental tension, i.e., how to learn representations that are both locally discriminative and globally coherent.

Several prior works have attempted to apply self-supervised learning (SSL) methods that have proven effective in centralized settings, e.g., SimCLR(Chen et al., 2020), BYOL(Grill et al., 2020), and SimSiam(Chen & He, 2021), to client-side local training within FL frameworks. However, these approaches often rely on large batch sizes or extensive negative samples, which are not applicable in resource-constrained FL environments. Crucially, as client heterogeneity increases, straightforward extensions of centralized SSL methods to FL scenarios often results in degraded performance. Alternative strategies have been proposed, including aggregating models via knowledge distillation(Han et al., 2022), local clustering (Lubana et al., 2022), and promoting unified representation by constraining consistent client model updating(Liao et al., 2024). While these methods have made partial strides, they primarily focus on preventing local representation collapse and lack explicit mechanisms to enforce semantic consistency across clients, rendering them vulnerable to representation drift under data heterogeneity. No existing approach adequately addresses both representation collapse and semantic misalignment in a unified manner so far, highlighting the need for a more principled approach to semantic-aware federated representation learning.

To bridge this critical gap, we propose Federated Unsupervised Learning with Prototype Anchored Consensus (FedPAC), a framework that resolves the tension between local learning and global consistency through a set of globally shared, learnable prototypes. On the client-side, we introduce a dual-alignment learning objective. At the representation level, we leverage self-supervised learning to promote discriminative feature learning and preventing collapse. At the semantic level, each client aligns local features to the prototypes, ensuring that representations corresponding to the similar concept are mapped to a globally consistent representation space regardless of local data distribution, significantly mitigating representation drift. On the server-side, we design a prototype aggregation strategy that refines the global prototypes by integrating semantic insights from clients, ensuring the prototypes remain diverse and globally representative throughout training. Through the interplay of local dual-alignment and server aggregation, FedPAC learns a unified representation space that is both locally discriminative and globally coherent, overcoming the limitations of prior federated unsupervised learning methods.

In summary, in this work we propose FedPAC to to tackle the key challenges of representation collapse and counteracts semantic misalignment in federated unsupervised learning. The core of our approach is a prototype-based semantic anchoring mechanism that establishes a globally consistent feature space across clients under non-IID data. (1) We propose a synergistic architecture that combines a dual-alignment learning objective for clients' local unsupervised learning and a prototype aggregation strategy refining global prototypes on the server. (2) We provide a rigorous convergence analysis that theoretically establishes the stability and soundness of our proposed method. (3) Empirically, we validate FedPAC through extensive experiments on two benchmark datasets. The results show that our framework significantly outperforms state-of-the-art methods, confirming the practical effectiveness of our semantic alignment strategy.

2 Related work

2.1 Self-supervised learning

SSL has advanced rapidly in recent years, enabling the learning of transferable representations without manual annotation. Current SSL methods are broadly categorized into discriminative and predictive approaches. Discriminative methods learn representations by enforcing invariance at the instance or cluster level. While effective in centralized settings, these methods face significant challenges under the constraints of federated learning. Contrastive learning, e.g., SimCLR(Chen et al., 2020), MoCo(He et al., 2020), are hampered by their reliance on large batch sizes or substantial negative samples, which are impractical on resource-constrained clients. Non-contrastive bootstrap methods like BYOL(Grill et al., 2020) and SimSiam(Chen & He, 2021), which typically depend on batch statistics for normalization and stabilization, are sensitive to data heterogeneity and can exacerbate client drift. Similarly, clustering approaches such as DeepCluster(Caron et al., 2018) and SwAV(Caron et al., 2020) often impose equipartition constraints to prevent collapse, which is ineffective under the imbalanced class distributions of non-IID data. Predictive methods, which learn through reconstruction (He et al., 2022) or pretext tasks (Gidaris et al., 2018), are less suitable for federated settings due to their high computational and communication costs.

2.2 FEDERATED UNSUPERVISED LEARNING

Recent research has begun to address the challenges of federated unsupervised learning, focusing on mitigating data heterogeneity and learning consistent representations across clients. A common strategy combines local SSL method with specialized aggregation mechanism. For instance, FedCA(Zhang et al., 2023) employs a shared dictionary to aggregate local representations and maintain the consistency of representation space, while ProtoFL(Kim et al., 2023) utilizes prototypical distillation to enhance global representations. FedU(Zhuang et al., 2021) aggregates only the online encoder parameters and incorporates a predictor adaptation module based on the divergence caused by non-IID data. Similarly, FedEMA(Zhuang et al., 2022) and FedX(Han et al., 2022) employ adaptive EMA decay based on local-global model divergence and bidirectional knowledge distillation, respectively, to jointly optimize both local and global models. Despite these advances, such methods often struggle under highly non-IID conditions and may raise privacy concerns. Beyond these, Orchestra(Lubana et al., 2022) uses local clustering tasks to learn representations and coordinates them through a hierarchical structure to enforce globally consistent cluster assignments. Recently, FedU2(Liao et al., 2024) alleviates collapse by encouraging uniform distribution of local representations and promotes uniformity by constraining consistent client model updates. While effectively mitigating representation collapse, they lack explicit mechanisms to ensure semantic consistency across clients, leading to persistent representation drift.

2.3 OPTIMAL TRANSPORT

Optimal Transport (OT)(Villani et al., 2008) is a mathematical framework for quantifying the discrepancies between probability distributions by seeking a probabilistic coupling that minimizes the total cost of transporting mass from one distribution to another. This formulation provides a mechanism for enforcing structural alignment between sets of elements such as representations. Consequently, OT has been widely studied and applied in machine learning for tasks like domain adaptation, robust clustering, and generative modeling(Courty et al., 2017; Tolstikhin et al., 2017). In representation learning, SwAV (Caron et al., 2020) leverages an OT-based assignment to perform online clustering and enable self-supervised learning without contrastive pairs, where entropic regularization helps prevent degenerate solutions and allows efficient optimization through Sinkhorn iterations.

3 PRELIMINARIES

Before detailing our methods, we introduce necessary definitions here. We also present Table 3 in Appendix A.2, providing a comprehensive explanation of the notations used throughout this paper.

We formulate Federated Unsupervised Learning (FUL) as follows. Consider a federated system consisting of a central server and N clients. Each client n holds a local unlabeled dataset $D_n = \{x_{n,i}\}_{i=1}^{|D_n|}$, where $|D_n|$ is the number of samples on client n. In practice, the data distributions across different clients are non-IID. A standard FUL problem can be formulated as a distributed optimization, aiming to collaboratively learn a global model θ across *N* clients, i.e.,

$$\min_{\theta} F(\theta) \;=\; \sum_{n=1}^{N} w_n \, F_n(\theta), \quad \text{with} \quad F_n(\theta) = \mathbb{E}_{x \sim D_n}\big[\mathcal{L}_n(\theta;\, x)\big], \tag{1}$$

where w_n denotes the aggregation weight of client n satisfying $w_n \geq 0$ and $\sum_n w_n = 1$, and $F_n(\theta)$ is the local objective based on client-specific unsupervised loss $L_n(\cdot)$.

For local unsupervised training we adopt the online-target network design similar to BYOL(Grill et al., 2020). The online network, parameterized by θ, decomposes into an encoder f_θ, a projector h_θ and a predictor q_θ. For an input x the online network outputs representation $\mathbf{y} = f_\theta(x)$, projection $\mathbf{z} = g_\theta(\mathbf{y})$ and prediction $q_\theta(\mathbf{z})$. The target network, parameterized by ϕ, is an exponential moving average of the online encoder and projector that receives no gradients and provides stable targets for the online predictor to match. Our method also relies on learning invariance to data augmentations. We define a stochastic augmentation function $\mathcal{T}$ that transforms an input x by randomly sampling $t' \sim \mathcal{T}$ to produce the augmented view $x' = t(x)$. In our setting, each client applies the same augmentation pipeline $\mathcal{T}$ independently to ensure consistency of augmented views across clients.

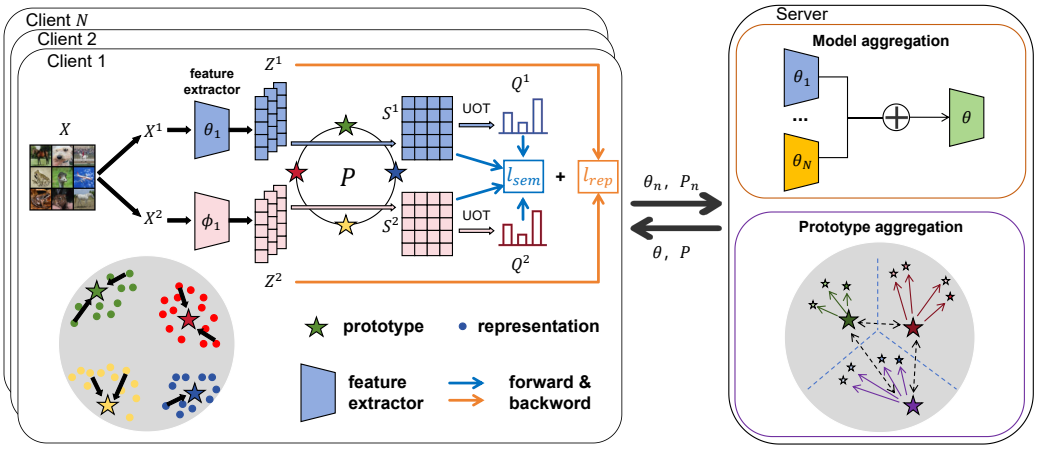


Figure 1: Overview of the FedPAC framework. clients perform local learning via a dual-alignment objective: a representation alignment loss enforces consistency between augmented views, while a semantic alignment loss that pulls representations toward their assigned prototypes. The server aggregates client models using Federated Averaging, and refines the global prototypes using an optimization-based aggregation strategy. This strategy steers each global prototype toward the centroid of local prototypes assigned to it, while preserving separation between each other, thereby strengthening global semantic consensus for the next round.

4 METHOD

4.1 OVERVIEW

To address the challenges of representation learning in federated unsupervised learning under non-IID data, we propose FedPAC. Our framework introduces two complementary objectives into the client-side unsupervised training, i.e., (1) a representation alignment loss that promotes the learning of view-invariant and discriminative features, thereby ensuring robustness against collapse, and (2) a semantic alignment loss that aligns local representations with a set of globally shared prototypes, thus enforcing a consistent semantic structure across the entire federation. This dual-pronged strategy yields a unified feature space that is both locally discriminative and globally consistent.

The overall training pipeline of FedPAC is illustrated in Figure 1. At the beginning of each communication round, the server broadcasts the current global model parameters and the global prototypes to the participating clients. Each selected client then perform E epochs of local unsupervised training by minimizing the local objective $\mathcal{L}_{\text{local}}$, which simultaneously optimizes its model parameters and local prototypes. Upon completion of local training, clients send their updated model parameters and local prototypes to the server for aggregation. Through iterating this process, the global model and prototypes are jointly refined, learning a effective and consistent representation space from distributed, unlabeled, and imbalanced data.

4.2 CLIENT-SIDE UNSUPERVISED TRAINING

When performing local unsupervised training, each client is designed to learn representations that are both discriminative and semantically consistent. For each batch, two augmented views of each sample are generated and processed by the feature extractor. We then compute the prototype assignments via optimal transport, predicated on the the similarity between projections and prototypes, and yields the corresponding semantic alignment loss ℓ_{sem}. Concurrently, a representation-alignment loss ℓ_{rep} is employed to align representations between views, supplemented by a rotation prediction task to enhance stability in early training. The comprehensive local objective is formulated as $\mathcal{L}_{\text{local}} = \ell_{\text{sem}} + \lambda\, \ell_{\text{rep}}$, minimized via SGD to jointly optimize model parameters and prototypes, with λ balancing the two terms. We detail the design of these two loss components below.

4.2.1 PROTOTYPE-BASED SEMANTIC ALIGNMENT

To facilitate semantic alignment across clients, we introduce a set of learnable prototypes $\{\mathbf{p}_k\}_{k=1}^{K}$ that serve as shared semantic anchors. Rather than directly aligning raw feature spaces, clients learn by mapping their local representations to a probability distribution over these shared prototypes. By enforcing consistency in these distributions for similar semantics across different views and clients, we achieve cross-view and cross-client semantic alignment anchored by the global prototypes.

Computing prototype assignments under heterogeneity. Given a batch of projections $\mathbf{Z} \in \mathbb{R}^{B\times d}$ and the prototypes $\mathbf{P} \in \mathbb{R}^{K\times d}$, our objective is to compute the prototype assignment matrix $\mathbf{Q} \in \mathbb{R}^{B\times K}$, where $\mathbf{Q}_{i,k}$ represents the probability mass assigned from sample sample i to prototype k. This both mitigates trivial collapse that arises from assigning each sample to its nearest prototype and provides a smooth training target for the cross-view prediction loss. We obtain the optimal $\mathbf{Q}$ by solving an optimal transport problem. In heterogeneous settings, clients often hold highly imbalanced data distributions. While enforcing a uniform distribution over prototype selections, as done in some clustering methods, is effective at preventing collapse, it can lead to inappropriate assignments and consequently degrade the quality of the learned representation. To address this challenge, we employ Unbalanced Optimal Transport (UOT) to compute the soft assignments. UOT allows the marginals to deviate from a strictly uniform distribution, thereby better accommodating scenarios where clients lack certain classes. In our setting we employ a UOT variant that replaces hard marginal equalities with KL penalties while retaining an entropic regularizer, i.e.,

$$\min_{\mathbf{Q}\in\mathbb{R}_{+}^{B\times K}} \operatorname{Tr}\big(-\mathbf{Q}^{\top}\mathbf{Z}\mathbf{P}^{\top}\big) - \varepsilon H(\mathbf{Q}) + \rho\,\mathrm{KL}\big(\mathbf{Q}^{\top}\mathbf{1}_B \,||\, \frac{1}{K}\mathbf{1}_K\big) \quad s.t.\ \mathbf{Q}\mathbf{1}_K = \frac{1}{B}\mathbf{1}_B, \tag{2}$$

which can be efficiently solved using Sinkhorn iterations. Here, $H(\mathbf{Q}) = -\sum_{i,k}\mathbf{Q}_{i,k}$ and the entropic parameter ε controls the smoothness of the assignment. $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence penalizing the deviation of prototype marginal $\mathbf{Q}^{\top}\mathbf{1}_B \in \mathbb{R}^K$ from uniform distribution. A hard uniform constraint is enforced on sample marginal $\mathbf{Q}\mathbf{1}_K$ to ensure equal assignment across samples, while the KL penalty with strength parameter $\rho \geq 0$ softly encourages balanced prototype usage. This asymmetric constraint design prevents trivial collapse while allowing prototype marginals to deviate from strict uniformity, thereby producing more reliable assignments under non-IID data and still promoting balanced prototype utilization. We can solve the above formula by using a sinkhorn-like iteration.

Swap loss with global prototypes. For each batch of samples, we generate two augmented views, yielding projections $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. We first compute the optimal soft assignment $\mathbf{Q}^{(1)}$ from $\mathbf{Z}^{(1)}$ using the above procedure, and similarly $\mathbf{Q}^{(2)}$ from $\mathbf{Z}^{(2)}$. Then we require the projections obtained from one view to predict the assignment of the other. Since projections lack interaction with the prototypes, we compute a similarity probability matrix and the semantic alignment loss is then defined as as the sum of cross-entropy between the assignment and similarity in both directions, i.e,

$$\ell_{\text{sem}} = -\frac{1}{2B}\sum_{i=1}^{B}\sum_{k=1}^{K}\Big(\mathbf{Q}_{i,k}^{(1)}\,\log\mathbf{S}_{i,k}^{(2)} + \mathbf{Q}_{i,k}^{(2)}\,\log\mathbf{S}_{i,k}^{(1)}\Big), \tag{3}$$

where $\mathbf{S}_{i,k} = \frac{\exp(\mathbf{z}_i^{\top}\mathbf{p}_k/\tau)}{\sum_{k'}\exp(\mathbf{z}_i^{\top}\mathbf{p}_{k'}/\tau)}$ with τ controlling the sharpness. Minimizing this loss encourages projections from different views of the same image to share the same prototype assignment. This objective jointly optimizes both representations and prototypes, simultaneously pulling each feature towards its assigned prototypes while also moving each prototype towards the centroid of its assigned features. Through this co-optimization, model structures the feature space into semantically distinct clusters anchored by prototypes, thereby promoting both view invariance and cross-client semantic alignment even under significant data heterogeneity.

4.2.2 CONTRASTIVE REPRESENTATION ALIGNMENT AND STABILIZATION

In federated unsupervised learning, the lack of ground truth labels often leads to weak supervisory signals, increasing the risk of representation collapse. Self-supervised learning mitigates this by en-

forcing augmentation invariance to promote compact and well-separated feature clusters. Integrated with our proposed semantic alignment loss, SSL methods not only align representations across different views to stabilize local training, but also yield discriminative features that facilitate semantic alignment. In our work, we adopt the similar architecture as BYOL for its stable and negative-free learning signal, which is well suited to resource-constrained FL environments. However, during the initial stages of training, prototypes may correspond to random or weakly discriminative features and thus provide unstable supervisory signals. Inspired by (Lubana et al., 2022) and other works that add inexpensive predictive tasks to stabilize early training, we introduce a rotation prediction task to encourage the formation of stable and meaningful representations before prototype stabilization.

Concretely, for each sample x we generate two augmented views $x^{(1)}, x^{(2)} \sim \mathcal{T}(x)$ along with rotated versions $\tilde{x}$. Rotation angles are randomly sampled from $\{0°, 90°, 180°, 270°\}$, with corresponding label $\varphi \in \{0, 1, 2, 3\}$. Following (Grill et al., 2020), we use an online and a target network to extract features from different views of the same sample, then a predictor maps the online projection to align with the target projection. Meanwhile, we attach a linear classification head $\omega \in \mathbb{R}^{d\times 4}$ to the online projection of $\tilde{x}$ to predict its rotation label, trained via the cross-entropy between the output logits and the true label φ. The representation alignment loss is therefore formulated as

$$\boldsymbol{\ell}_{\text{rep}} = \frac{1}{2B}\sum_{i=1}^{B}\sum_{v=1}^{2}\Big(\|q_\theta(\mathbf{z}_i^{(v)}) - \bar{\mathbf{z}}_i^{(v')}\|_2^2 + \mathrm{CE}\big(\omega(\tilde{\mathbf{z}}_i^{(v)}),\ \varphi_i\big)\Big), \tag{4}$$

where $v' = 3 - v$ denotes the other augmented view, $\bar{\mathbf{z}}$ denotes the projection outputted by target network with stop-gradient applied, and $\mathrm{CE}(\cdot,\cdot)$ is the cross-entropy function. This combined objective encourages the online network to learn view-invariant and discriminative features, while the rotation prediction task provides an additional supervisory signal to mitigate collapse in early stage.

4.3 Server-side model aggregation

In each communication round, once participating clients complete local training they upload both model parameters and local prototypes to the server. Model parameters are aggregated via weighted averaging, using each client's number of local samples as the aggregation weight. Local prototypes represent semantic centers specific to local data distribution, simply averaging would blur these distinct semantic clusters and destroy the learned structure. Therefore, we propose an optimization-based aggregation aggregation mechanism designed to consolidate local prototypes into an updated set of global prototypes, thereby preserving a coherent cross-client semantic structure.

Let the current global prototypes $\mathbf{P}_g \in \mathbb{R}^{K\times d}$ serve as fixed semantic anchors, and the collection of all local prototypes from participating clients $\mathbf{P}_l \in \mathbb{R}^{I\times d}$ (typically $I > K$) be the set to be assigned over anchors. We compute a similarity matrix $\mathbf{S}$ between $\mathbf{P}_l$ and $\mathbf{P}_g$, and a soft assignment matrix $\mathbf{Q}$ that indicates how strongly each local prototype is assigned to each global prototype via the balanced version of equation 2, i.e., with strict equipartition constraints on both marginals. We formulate the aggregation of local prototypes as an optimization objective with two complementary losses. The first term, an assignment fidelity loss, is formulated as the cross-entropy between the assignment and similarity:

$$\boldsymbol{\ell}_{\text{fed}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}\mathbf{Q}_{i,k}\log\mathbf{S}_{i,k}, \tag{5}$$

which encourages consistency between the similarity and the soft assignment. Minimizing $\mathcal{L}_{\text{fed}}$ pulls each global prototype towards the centroid of local prototypes that are strongly assigned to it, ensuring the updated global prototypes reflect the consensus of the local semantic centers. And the second term, a prototype uniformity loss that penalizes excessive proximity between prototypes via pairwise repulsion, serves as as a regularizer to encourage the updated global prototypes to remain well separated. This prevents the prototypes from collapsing into a few redundant clusters and preserves the overall semantic diversity. It is defined as

$$\boldsymbol{\ell}_{\text{uni}} = \log\left(\frac{1}{I(I-1)}\sum_{i\neq j}\exp\big(-2\gamma\,\|\mathbf{p}_i - \mathbf{p}_j\|^2\big)\right), \tag{6}$$

where $\gamma > 0$ controls the sharpness of the exponential weighting. The server updates the global prototypes by minimizing the combined objective $\mathcal{L}_{\text{proto}} = \boldsymbol{\ell}_{\text{fed}} + \beta\, \boldsymbol{\ell}_{\text{uni}}$, where β is a trade-off coefficient. The composite objective encourages global prototypes to align with the representative local semantics while preserving sufficient separation, thereby capturing cross-client semantic structure and preventing collapse. After aggregation, the updated global prototypes are distributed to clients for next round of training, providing a refined and globally consistent semantic guidance.

5 Convergence Analysis

In this section, we provide a convergence analysis for our proposed federated unsupervised learning framework. We first detail the assumptions that underpin our analysis and then present our main theorems: one characterizing the sufficient decrease achieved by our prototype aggregation method and the other guaranteeing the global convergence rate of the entire algorithm. Our analysis quantitatively demonstrates how factors influence the final solution quality and precise bounds are provided in theorems below.

To facilitate our theoretical analysis, we introduce the following notation. Let θ denote the model parameters and ρ denote the prototype parameters, and both are optimized jointly during local training. Thus, we define the combined parameter $\psi = (\theta, \rho)$. Any assumption stated for ψ is understood to hold for both θ and ρ. Clients perform local updates with learning rate η_l, while the server performs U steps of SGD with learning rate η_ρ on the surrogate objective $F_s(\cdot)$. For brevity, detailed statements of the assumptions and the complete convergence proof are provided in Appendix B.

Theorem 1 (Sufficient Decrease of Prototype Aggregation)**.** *Let Assumption 1 and 6 holds, and the server-side learning rate satisfies $\eta_\rho \leq 1/L$, then the proposed prototype optimization step yields*

$$\mathbb{E}[F(\theta_r, \rho_{r+1})] \leq F(\theta_r, \rho_r) - \frac{\eta_\rho}{2} \sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho U \zeta^2}{2}.$$

This theorem formally bridges the gap between the server's surrogate optimization task and the true global objective. It guarantees that our prototype aggregation strategy achieves a sufficient decrease in the global loss each round, thus providing a principled convergence guarantee even when the server operates with limited information based on not exactly the same objective.

Theorem 2 (Global Convergence)**.** *Let Assumptions 1-6 hold and the server-side learning rate satisfies $\eta_\rho \leq 1/L$. After R communication rounds, the average expected squared norm of the global gradient is bounded as follows:*

$$\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\|\nabla F(\psi_r)\|^2 \leq \frac{F(\psi_0) - F^\star}{R\,\Gamma_1} + \frac{\Gamma_2}{\Gamma_1},$$

where F^* is the minimum value of the global objective, $\Gamma_1 = \min(\frac{\eta_l E M}{2N}, \frac{\eta_\rho}{2})$ and $\Gamma_2 = \frac{G^2 M}{N}\left(\frac{L^2\eta_l^2 E^2}{2} + \frac{\eta_l^3 E^3 L^2}{3}\right) + \left(\eta_l + \frac{L^2\eta_l^2 E M}{2N} + \frac{\eta_l^3 E^2 L^2 M}{2N}\right)\bar{\sigma}^2 + \frac{\eta_\rho U \zeta^2}{2}$.

This theorem demonstrates that our algorithm converges to a neighborhood of a stationary point at a sublinear rate of $\mathcal{O}(\frac{1}{R})$. Specifically, the size of this neighborhood is influenced by the state of initial model, data heterogeneity across clients, the number of epochs, learning rates and the error from our prototype aggregation scheme, quantifying the trade-offs inherent in the federated learning.

6 Experiments

6.1 Experimental setup

Datasets and partitioning. We conduct experiments on CIFAR-10 and CIFAR-100(Krizhevsky et al., 2009). To simulate non-IID data distributions across clients, we partition the training set among N clients using a Dirichlet distribution(Hsu et al., 2019). Specifically, for each class we sample a probability vector from $\text{Dir}(\alpha)$ and distribute samples to clients accordingly. The concentration parameter α controls the heterogeneity, with a smaller α results in more skewed partitions.

Table 1: Accuracy (%) on CIFAR10 and CIFAR100 under non-IID ($\alpha = 0.1$) cross-device (100 clients) and cross-silo (10 clients) settings. FedU2 can be combined with different SSL methods and we list the results for all of them(denoted by superscripts). Evaluation is performed via linear probing and semi-supervised fine-tuning with 1%/10% labelled data.

Dataset	CIFAR-10						CIFAR-100					
Setting	Cross-Device (N=100)			Cross-Silo (N=10)			Cross-Device (N=100)			Cross-Silo (N=10)		
Method	Linear	1%	10%	Linear	1%	10%	Linear	1%	10%	Linear	1%	10%
SimCLR	61.70	48.76	68.82	74.63	63.34	76.25	34.42	13.25	36.61	50.42	19.25	43.61
BYOL	60.91	51.46	68.1	75.52	75.55	81.93	29.74	11.56	33.13	48.87	20.52	43.15
SimSiam	63.63	54.81	71.14	78.51	70.32	78.33	32.96	12.71	35.9	50.38	23.71	43.96
FedU	59.72	50.59	72.42	80.03	69.51	83.15	31.74	12.09	34.08	54.36	30.97	47.46
FedX	68.38	58.55	73.59	74.93	67.47	81.38	35.78	15.91	36.3	49.02	19.57	41.69
FedEMA	63.8	53.21	73.56	78.95	68.65	81.87	32.49	12.95	36.29	51.66	32.35	47.21
Orchestra	64.28	53.69	69.32	76.13	75.8	**85.7**	27.66	11.95	33.21	52.81	33.7	48.89
FedU2$^{\text{SimCLR}}$	65.58	54.5	72.44	80.43	73.47	82.66	36.07	16.25	35.77	53.55	31.98	48.71
FedU2$^{\text{SimSiam}}$	69.01	60.71	73.97	82.19	73.6	82.48	34.18	12.81	34.25	54.05	30.57	50.93
FedU2$^{\text{BYOL}}$	68.81	56.69	72.56	82.8	75.04	84.67	36.25	17.01	**38.0**	54.78	32.19	**52.69**
FedPAC	**71.36**	**62.17**	**75.03**	**83.56**	**77.59**	84.47	**37.32**	**17.68**	37.64	**56.33**	**37.23**	51.98

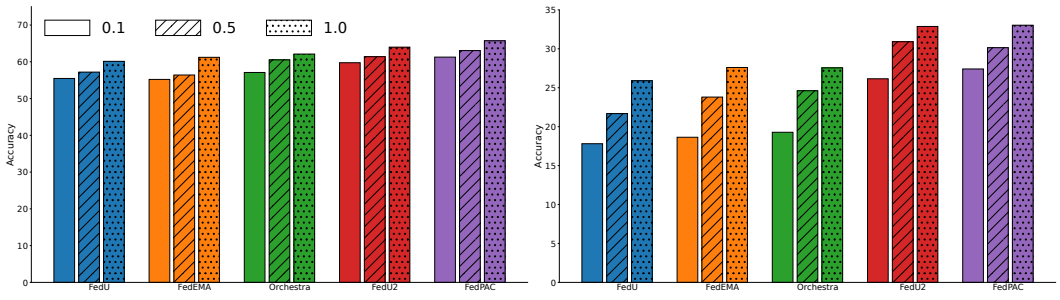
Evaluation Protocol. We evaluate representation quality using linear probing(Chen et al., 2020), K-nearest neighbors (KNN) classification(Chen & He, 2021) and semi-supervised fine-tuning. Linear probing trains a linear classifier on frozen features to assess the linear separability of the representations, while KNN classification on frozen embeddings provides a label-free measure of feature discriminability. Semi-supervised fine-tuning with 1% or 10% labeled data evaluates representation transferability in low-label regimes.

Comparative Methods. We compare **FedPAC** with several relevant baselines: (1) federated adaptations of centralized self-supervised learning methods, including SimCLR, BYOL, and SimSiam, combined with FedAvg, (2) the state-of-the-art federated unsupervised learning methods including FedU(Zhuang et al., 2021), FedX(Han et al., 2022), FedEMA(Zhuang et al., 2022), Orchestra(Lubana et al., 2022), and FedU2(Liao et al., 2024). For fair comparison, all methods use the same encoder architecture and data partitions. Implementations follow the original papers and, where available, rely on official codebases. Results are averaged over 3 independent runs with different random seeds.

6.2 Experiment Results

Representation Evaluation. Following existing methods(Liao et al., 2024; Lubana et al., 2022), we first assess the quality of the learned representations via linear probing and semi-supervised fine-tuning. The comprehensive results are presented in Table 1 and key observations are summarized as follows. **First**, simply combining centralized SSL methods with FedAvg yields limited accuracy under high data heterogeneity, confirming their fragility to non-IID data. **Second**, FedPAC consistently outperforms all baselines in linear evaluation across both cross-silo and cross-device settings, demonstrating the superior quality of learned representations. **Third**, under the challenging 1% semi-supervised setting, FedPAC achieves a substantial performance margin over compared methods. This can be attributed to its well-structured feature space, which already exhibits distinct and semantically coherent clusters. Consequently, downstream adaptation task is simplified to learning a linear mapping from these clusters to their corresponding labels, requiring minimal supervision. While this gap narrows as the proportion of labeled data increases to 10%, FedPAC remains highly competitive. The superior performance of FedPAC reported above underscores its ability to learn discriminative and transferable representations.

Analysis of Sensitivity to hyper-parameters. We begin by evaluating the sensitivity of different methods to data heterogeneity, controlling $\alpha = \{0.1, 0.5, 1.0\}$. Results in Figure 2 show that FedPAC maintains stable performance even under severe non-IID settings, whereas the accuracy of baseline methods degrades, particularly on the more complex CIFAR-100 dataset. This demon-



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Sensitivity to data heterogeneity on CIFAR-10 (left) and CIFAR-100 (right). FedPAC is more robust to data heterogeneity.

strates the robustness of our prototype-anchored consensus against skewed data distributions. For additional experiments analyzing the effects of other parameters, please refer to the Appendix C.

Ablation Study We conduct an ablation study under cross-device setting to evaluate the contribution of each component in FedPAC. As shown in Table 2, both the semantic alignment loss and the representation alignment loss are essential for achieving optimal performance. The semantic alignment loss contributes the most to cross-client consistency, while the representation alignment loss is critical in preventing representation collapse. The prototype aggregation strategy is also necessary for maintaining global semantic consensus. Removing any component leads to performance degradation, which validates our design choices. We also plot the convergence curves of FedPAC and FedU2 in Figure 3. Compared to FedU2, our method exhibits faster convergence and a more stable training process with less oscillations throughout. In contrast, the varian of FedPAC without ℓ_{rep} shows slower accuracy improvement during early training and suffers from instability, further confirming its importance in improving representation quality and stabilizing training.

Table 2: Comparison of FedPAC and its ablated variants on CIFAR10 and CIFAR100 under non-IID ($\alpha = 0.1$) cross-device settings.

Dataset	Method	FedPAC	w/o ℓ_{sem}	w/o ℓ_{rep}	w/o $\mathcal{L}_{\text{proto}}$
CIFAR-10	KNN	63.99	55.91	59.87	61.48
	Linear	71.36	62.55	68.4	69.22
	1%	62.17	53.36	60.4	60.62
	10%	75.03	71.37	73.55	73.09
CIFAR-100	KNN	27.41	19.68	20.16	23.02
	Linear	37.32	30.4	30.33	33.54
	1%	17.68	12.99	12.01	14.01
	10%	37.64	30.28	30.86	31.45

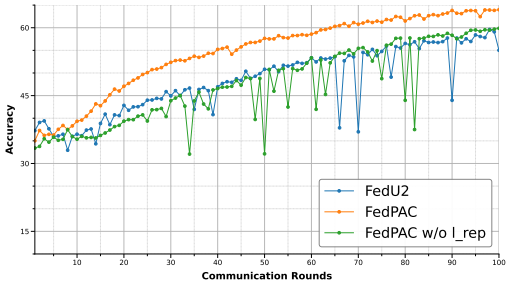


Figure 3: Convergence curve of KNN accuracy versus communication rounds on CIFAR-10.

7 Conclusion

In this work, we propose FedPAC to address the critical challenges of representation collapse and semantic misalignment in federated unsupervised learning under non-IID data. It leverages prototypes as semantic anchors to establish a semantic consensus among clients, enables learning discriminative and semantically consistent representations from distributed, unlabeled, and imbalanced data. Theoretically we provides a rigorous convergence analysis, and empirically, we conduct experiments on CIFAR10 and CIFAR100 to validate the superior performance of FedPAC.

References

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer*

vision (ECCV), pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pp. 691–707. Springer, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Hansol Kim, Youngjun Kwak, Minyoung Jung, Jinho Shin, Youngsung Kim, and Changick Kim. Protofl: Unsupervised federated learning via prototypical distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6470–6479, 2023.

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

Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. Rethinking the representation in federated unsupervised learning with non-iid data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22841–22850, 2024.

Ekdeep Lubana, Chi Ian Tang, Fahim Kawsar, Robert Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14461–14484. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/lubana22a.html`.

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

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Fengda Zhang, Kun Kuang, Long Chen, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Fei Wu, Yueting Zhuang, et al. Federated unsupervised representation learning. *Frontiers of Information Technology & Electronic Engineering*, 24(8):1181–1193, 2023.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4912–4921, 2021.

Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022.

A Method Details

A.1 Computing prototype assignments

We now describe the process of computing prototype assignments in more detail. Given batch projections $\mathbf{Z} \in \mathbb{R}^{B\times d}$ and the prototypes $\mathbf{P} \in \mathbb{R}^{K\times d}$, we seek the prototype assignment matrix $\mathbf{Q} \in \mathbb{R}^{B\times K}$ where $\mathbf{Q}_{i,k}$ denotes the mass that sample i places on prototype k. Such a soft coupling both mitigates trivial collapse that arises from assigning each sample to its nearest prototype and supplies a smooth training target for cross-view prediction losses. We obtain the optimal $\mathbf{Q}$ by solving an optimal transport problem. In heterogeneous settings, clients may hold very different numbers and categories of samples, or lack certain categories. Strictly enforcing even selection of samples and prototypes, although effective at preventing collapse, can force inappropriate assignments that reduce fidelity. Unbalanced optimal transport (UOT) relaxes marginal equalities while penalizing the deviation via a divergence and is more suitable for solving this problem. In our setting we employ a UOT variant that replaces hard marginal equalities with KL penalties while retaining an entropic regularizer, i.e.,

$$\min_{\mathbf{Q}\in\mathbb{R}_+^{B\times K}} \langle \mathbf{Q}, \mathbf{C}\rangle - \varepsilon H(\mathbf{Q}) + \rho\,\mathrm{KL}\big(\mathbf{Q}^\top \mathbf{1}_B \,||\, \mathbf{a}\big) + \mu\,\mathrm{KL}\big(\mathbf{Q}\mathbf{1}_K \,||\, \mathbf{b}\big), \tag{7}$$

where cost matrix $\mathbf{C} = -\mathbf{Z}\mathbf{P}^\top$. Here $\mathbf{Q}\mathbf{1}_K \in \mathbb{R}^B$ and $\mathbf{Q}^\top\mathbf{1}_B \in \mathbb{R}^K$ are the sample and prototype side marginals respectively, while $\mathbf{a} \in \mathbb{R}^B$ and $\mathbf{b} \in \mathbb{R}^K$ are their target margins. $\rho, \mu \geq 0$ weight the marginal matching strength on the sample and prototype sides. To address the heterogeneous setting, we preserve a hard marginal on the sample side to ensure that each sample is equally assigned. This is implemented by taking $\mu \to \infty$ and hence imposing $Q\mathbf{1}_K = \mathbf{a} = \frac{1}{B}\mathbf{1}_B$. And we relax the prototype-side marginal via a KL penalty with target $\mathbf{b} = \frac{1}{K}\mathbf{1}_K$, we can rewrite equation 7 as:

$$\min_{\mathbf{Q}\in\mathbb{R}_+^{B\times K}} \mathrm{Tr}\big(-\mathbf{Q}^\top\mathbf{Z}\mathbf{P}^\top\big) - \varepsilon H(\mathbf{Q}) + \rho\,\mathrm{KL}\big(\mathbf{Q}^\top\mathbf{1}_B \,||\, \frac{1}{K}\mathbf{1}_K\big) \quad s.t.\ \mathbf{Q}\mathbf{1}_K = \frac{1}{B}\mathbf{1}_B. \tag{8}$$

Compared with strict equipartition, the one-sided unbalanced marginal constraint prevents trivial collapse while allowing prototype marginals to deviate from exact uniformity. This flexibility avoids spurious assignments when clients lack certain classes, produces more reliable assignments under non-IID data and still encourages balanced prototype utilization.

With entropy regularization and KL relaxation, the solution of equation 2 admits a Gibbs-like factorization $\mathbf{Q}^\star = \mathrm{diag}(\mathbf{u})\ \mathbf{G}\ \mathrm{diag}(\mathbf{v})$ with the Gibbs kernel matrix $\mathbf{G} = \exp(\mathbf{Z}\mathbf{P}^\top/\varepsilon)$. For numerical stability, we compute the multiplicative renormalizers $\mathbf{u}$, $\mathbf{v}$ by iterating the following calculations in the logarithmic domain:

$$\log \mathbf{u} \leftarrow \log \mathbf{a} - \log(\mathbf{G}\mathbf{v}), \qquad \log \mathbf{v} \leftarrow \kappa\big(\log \mathbf{b} - \log(\mathbf{G}^\top \mathbf{u})\big), \tag{9}$$

where $\kappa = \rho/(\rho + \varepsilon)$. After convergence we plug $\mathbf{u}$ and $\mathbf{v}$ back into the original factorization to obtain $\mathbf{Q}^\star$. Optionally one may obtain a discrete assignment from $\mathbf{Q}^\star$ by a rounding procedure, but we retain the continuous soft assignment in training because it provides smoother gradients and better numerical stability.

Working on small batches. When the batch size B is much smaller than the number of prototypes K, an equal partitioning of the batch samples across K prototypes is infeasible. To mitigate this issue, we augment the current batch with a memory queue containing the projections from recent training samples. We solve the UOT problem over this augmented set and only the entries of the assignment matrix corresponding to the samples from the current batch are utilized to compute the semantic alignment loss. This memory-augmented strategy enables the estimation of stable assignments with small batch sizes, while imposing minimal computational and memory overhead.

A.2 Notation and Algorithm

We present Table 3 to better summarize and explain the notations used in this paper. And we also summarize the entire framework in Algorithm 1 that better illustrates the entire training process.

Notation	Explanation
R, r	Total number of communication rounds, current round
N, n	Total number of clients, local client index
θ_r	Global model parameters at r-th round
$f_\theta,\ g_\theta,\ q_\theta$	Online encoder, online projector and predictor parameterized by θ
$f_\phi,\ g_\phi$	Target encoder and target projector parameterized by ϕ
$\mathbf{z}$	Representation vector (output of encoder)
D_n	Local dataset of client n
$x_{n,i}$	The i-th sample of client n (original input)
$\mathcal{T}$	Stochastic augmentation function
$\tilde{x} = t(x),\ t \sim \mathcal{T}$	Apply a random data transform to x to get the augmented view
E	The number of epochs for local training on each client
B	Batch size for local training
K	Total number of global prototypes
$\mathbf{P}$, $\mathbf{p}_k$	Global prototype matrix, k-th prototype vector
d	Dimensions of representation and prototype vectors
$\mathbf{Z}$	The representation matrix of a batch
$\mathbf{S}$	Similarity probability matrix
$\mathbf{Q}$	Prototype assignments
I	Number of local prototypes of the clients participating in the training
ρ	Parameterized representation of prototype vectors
$\psi = (\theta, \rho)$	Joint parameterized representation of models and prototypes
U	The number of epochs for optimizing global prototypes on the server
$F_s(\cdot)$	Server-side surrogate function for optimizing global prototypes
η_l, η_ρ	Learning rate for clients' local training, and server-side optimization

Table 3: Caption

A.3 Implementation details

We adopt ResNet-18(He et al., 2016) as the encoder architecture. Projector and predictor designs and EMA rules follow their original papers. Unless specified, the number of local epochs E to 10, the total communication rounds R to 100, and the Dirichlet parameter α to 0.1 to simulate high data heterogeneity. Experiments cover two FL setting: a cross-silo scenario with $N = 10$ clients and a participation rate 1.0, and a cross-device scenario with $N = 100$ clients and a participation rate 0.1. We use a batch size of 64 and 32 prototypes as semantic anchors for CIFAR-10, and a batch size of 128 and 128 prototypes for CIFAR-100.

Algorithm 1 The FedPAC Framework

Input: Number of clients N, communication round R, local training epochs on clients E, prototypes optimization epochs on server U, local data $D_n = \{x_{n,i}\}_{i=1}^{|D_n|}$, trade-off coefficient $\lambda,\ \beta$.
Output: Global model θ_R.
Server executes:
Initialize global model θ_0 and global prototypes $\mathbf{P}_g$;
for each communication round $r = 0, 1, \ldots, R-1$ **do**
    Randomly select a subset of clients S_r;
    for each client $n \in S_r$ **in parallel do**
        Send $\theta_r,\ \mathbf{P}_r$ to client n;
        Client executes($n,\ \theta_r,\ \mathbf{P}_r$);
        Aggregate models parameters: $\theta_{r+1} \leftarrow \text{FedAvg}(\theta_n|_{n \in S_r})$;
        for prototypes optimization epoch $u = 1, 2, \ldots, U$ **do**
            Compute assignment fidelity loss ℓ_{fed} using equation 5;
            Compute prototype uniformity loss ℓ_{uni} using equation 6;
            Update global prototypes $\mathbf{P}_g$ by minimizing $\mathcal{L}_{\text{proto}} = \ell_{\text{fed}} + \beta\, \ell_{\text{uni}}$;
        end for
    end for
end for
return θ_R
Client executes($n,\ \theta_r,\ \mathbf{P}_g$)**:**
$\theta_n \leftarrow \theta_r, \mathbf{P}_n \leftarrow \mathbf{P}_g$;
for each local epoch $e = 1, 2, \ldots, E$ **do**
    Sample a mini-batch form D_n;
    Compute semantic alignment loss ℓ_{sem} using equation 3;
    Compute representation alignment ℓ_{rep} using equation 4;
    Update local model θ_n and prototypes $\mathbf{P}_n$ by minimizing $\mathcal{L}_{\text{local}} = \ell_{\text{sem}} + \lambda\, \ell_{\text{rep}}$;
end for
return updated θ_n and $\mathbf{P}_n$ back to the server

B Proof of Theorems

Let θ denote the model parameters and ρ denote the prototype parameters, and both are optimized jointly during local training. Thus, we define the combined parameter $\psi = (\theta, \rho)$. Any assumption stated for ψ is understood to hold for both θ and ρ. Clients perform local updates with learning rate η_l, while the server performs U steps of SGD with learning rate η_ρ on the surrogate objective $H(\cdot)$.

B.1 Assumptions

We base our convergence analysis on the following standard assumptions.

Assumption 1 (Smoothness)**.** *Local objective functions $F_1,\ F_2,\ \ldots\ , F_N$ are all L-smooth, i.e., $\|\nabla F_n(\psi) - \nabla F_n(\psi')\| \le L\|\psi - \psi'\|$ for $n = 1, \ldots, N$.*

Assumption 2 (Unbiased Gradient)**.** *Let ξ denotes a batch of samples uniformly sampled at random from local data. The stochastic gradient is an unbiased estimator of the true local gradient, i.e., $\mathbb{E}\left[\nabla F_n(\psi, \xi)\right] = \nabla F_n(\psi)$.*

Assumption 3 (Bounded Variance)**.** *The variance of the stochastic gradient on each client n is bounded: $\mathbb{E}\|\nabla F_n(\psi, \xi) - \nabla F_n(\psi)\|^2 \le \sigma_n^2$ for $n = 1, \ldots, N$.*

Assumption 4 (Bounded Gradient Norm)**.** *The expected squared norm of any client's stochastic gradient is uniformly bounded: $\mathbb{E}\|\nabla F_n(\psi, \xi)\|^2 \le G^2,$ for $n = 1, \ldots, N$.*

Assumption 5 (Uniform Client Sampling)**.** *In each communication round r, a set S_r of M clients is selected uniformly at random from the total N clients. The probability of any client n being selected is $\mathbb{P}(n \in S_r) = M/N$.*

Assumption 6 (Bounded Prototype Gradient Estimation Error)**.** *The gradient of the server-side surrogate function $\nabla F_s(\rho_r)$ is an estimator of the true global gradient with bounded variance. Specifically, at communication round r, we have $\mathbb{E}\|\nabla F_s(\rho_r) - \nabla_\rho F(\theta_r, \rho_r)\|^2 \le \zeta^2$.*

B.2 Convergence Analysis

Let $\rho_0 = \rho_r$ and $\rho_U = \rho_{r+1}$ denote the initial and final prototype parameters at round r, respectively. The update rule for each step u is:

$$\rho_{u+1} = \rho_u - \eta_\rho \nabla F_s(\rho_u). \tag{10}$$

We first prove Theorem 1, which establishes that our prototype aggregation strategy guarantees a sufficient decrease in the global objective function.

Theorem 1 (Sufficient Decrease Guarantee of Prototype Aggregation)**.** *Let Assumption 1 and 6 holds, and the server-side learning rate satisfies $\eta_\rho \leq 1/L$, then the proposed prototype optimization step yields*

$$\mathbb{E}[F(\theta_r, \rho_{r+1})] \leq F(\theta_r, \rho_r) - \frac{\eta_\rho}{2} \sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho U \zeta^2}{2}.$$

Proof. With Assumption 1 holds, considering ψ_r as fixed during the server-side prototype aggregation, it follows that

$$F(\theta_r, \rho_u) \leq F(\theta_r, \rho_u) + \langle \nabla_\rho F(\theta_r, \rho_u), \rho_{u+1} - \rho_u \rangle + \frac{L}{2}\|\rho_{u+1} - \rho_u\|^2. \tag{11}$$

Then substituting the update difference $\rho_{u+1} - \rho_u = -\eta_\rho \nabla F_s(\rho_u)$ into equation 11, we have

$$F(\theta_r, \rho_{u+1}) \leq F(\theta_r, \rho_u) - \eta_\rho \langle \nabla_\rho F(\theta_r, \rho_u), \nabla F_s(\rho_u) \rangle + \frac{L\eta_\rho^2}{2}\|\nabla F_s(\rho_u)\|^2. \tag{12}$$

Taking the expectation on both sides of the above formula, we have

$$\mathbb{E}[F(\theta_r, \rho_{u+1})] \leq F(\theta_r, \rho_u) - \eta_\rho \underbrace{\mathbb{E}\langle \nabla_\rho F(\theta_r, \rho_u), \nabla F_s(\rho_u) \rangle}_{A_1} + \frac{L\eta_\rho^2}{2}\mathbb{E}\|\nabla F_s(\rho_u)\|^2. \tag{13}$$

Using the identity $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ and let $a = \nabla_\rho F(\theta_r, \rho_u)$ and $b = \nabla F_s(\rho_u)$, it follows that

$$\begin{aligned} A_1 &= \mathbb{E}\left[\frac{1}{2}\Big(\|\nabla_\rho F(\psi_r, \rho_u)\|^2 + \|\nabla F_s(\rho_u)\|^2 - \|\nabla_\rho F(\theta_r, \rho_u) - \nabla F_s(\rho_u)\|^2\Big)\right] \\ &= \frac{1}{2}\Big(\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \mathbb{E}\|\nabla F_s(\rho_u)\|^2 - \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u) - \nabla F_s(\rho_u)\|^2\Big). \end{aligned}$$

Plugging back into equation 13, we have

$$\begin{aligned} \mathbb{E}[F(\theta_r, \rho_u)] &\leq F(\theta_r, \rho_r) + \frac{L\eta_\rho^2}{2}\mathbb{E}\|\nabla F_s(\rho_u)\|^2 \\ &\quad - \frac{\eta_\rho}{2}\Big(\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \mathbb{E}\|\nabla F_s(\rho_u)\|^2 - \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u) - \nabla F_s(\rho_u)\|^2\Big) \\ &= F(\theta_r, \rho_r) - \frac{\eta_\rho}{2}\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 - \Big(\frac{\eta_\rho}{2} - \frac{L\eta_\rho^2}{2}\Big)\mathbb{E}\|\nabla F_s(\rho_u)\|^2 \\ &\quad + \frac{\eta_\rho}{2}\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u) - \nabla F_s(\rho_u)\|^2. \end{aligned} \tag{14}$$

We choose the server-side learning rate such that $\eta_\rho \leq 1/L$, which implies that the coefficient of the $\mathbb{E}\|\nabla F_s(\rho_u)\|^2$ term is non-negative. We can thus drop this term to obtain a valid upper bound:

$$\mathbb{E}[F(\theta_r, \rho_{u+1})] \leq F(\theta_r, \rho_u) - \frac{\eta_\rho}{2}\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho}{2}\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u) - \nabla F_s(\rho_u)\|^2. \tag{15}$$

Applying Assumption 6 to bound the last term, we have

$$\mathbb{E}[F(\theta_r, \rho_{u+1})] \leq F(\theta_r, \rho_u) - \frac{\eta_\rho}{2}\mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho \zeta^2}{2}. \tag{16}$$

Note that $\sum_{u=0}^{U-1}(\mathbb{E}\left[F(\rho_{u+1}) - F(\rho_r)\right] = \mathbb{E}[F(\rho_{u+1})] - F(\rho_r)$, and summing over $u = 0, \ldots, U-1$, we have

$$\mathbb{E}[F(\theta_r, \rho_{r+1})] \leq F(\theta_r, \rho_r) - \frac{\eta_\rho}{2} \sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho U \zeta^2}{2}. \tag{17}$$

This completes the proof.

□

Before presenting the Theorem 2, we state and prove several lemmas to clearly express our subsequent proof clearly.

Lemma 1 (Local Model Divergence)**.** *Let Assumptions 2 to 4 holds, the expected squared deviation between the the initial global model ψ_r and local model parameters ψ_n^E after E local update steps is bounded as follows that*

$$\mathbb{E}\left[\|\psi_n^E - \psi_r\|^2\right] \leq \eta_l^2 E^2 G^2 + \eta_l^2 E \sigma_n^2$$

Proof. Note that the total change in the local model parameters on client n after E steps is the sum of the individual updates with local learning rate η_l, we have

$$\psi_n^E - \psi_r = \sum_{e=0}^{E-1} (\psi_n^{e+1} - \psi_n^e) = -\eta_l \sum_{e=0}^{E-1} \nabla F_n(\psi_n^e, \xi_n^e). \tag{18}$$

Taking the expected squared norm, we have

$$\mathbb{E}\|\psi_n^E - \psi_r\|^2 = \eta_l^2 \mathbb{E} \left\| \sum_{e=0}^{E-1} \nabla F_n(\psi_n^e, \xi_n^e) \right\|^2. \tag{19}$$

To analyze the sum of gradients, we decompose each stochastic gradient into the true local gradient and a zero-mean noise term δ_n^e such that

$$\nabla F_n(\psi_n^e, \xi_n^e) = \nabla F_n(\psi_n^e) + \delta_n^e, \tag{20}$$

where $\delta_n^e = \nabla F_n(\psi_n^e, \xi_n^e) - \nabla F_n(\psi_n^e)$. Plugging it into equation 19, we have

$$\mathbb{E}\|\psi_n^E - \psi_r|^2 = \eta_l^2 \mathbb{E} \left\| \sum_{e=0}^{E-1} (\nabla F_n(\psi_n^e) + \delta_n^e) \right\|^2. \tag{21}$$

Then we expand the squared norm and get

$$\left\| \sum_{e=0}^{E-1} (\nabla F_n(\psi_n^e) + \delta_n^e) \right\|^2 = \left\| \sum_{e=0}^{E-1} \nabla F_n(\psi_n^e) \right\|^2 + \left\| \sum_{e=0}^{E-1} \delta_n^e \right\|^2 + 2 \left\langle \sum_{e=0}^{E-1} \nabla F_n(\psi_n^e), \sum_{e=0}^{E-1} \delta_n^e \right\rangle. \tag{22}$$

Note that the expectation of the cross-term is zero with Assumption 2 holds. Thus, we have

$$\mathbb{E}\|\psi_n^E - \psi_r\|^2 = \eta_l^2 \left(\mathbb{E} \left\| \sum_{e=0}^{E-1} \nabla F_n(\psi_n^e) \right\|^2 + \mathbb{E} \left\| \sum_{e=0}^{E-1} \delta_n^e \right\|^2 \right). \tag{23}$$

Since the noise terms δ_n^e are independent across steps and have zero mean conditioned on the history, their cross terms vanish in expectation. Therefore, we can rewrite the second term as

$$\begin{aligned} \mathbb{E} \left\| \sum_{e=0}^{E-1} \delta_n^e \right\|^2 &= \sum_{e=0}^{E-1} \mathbb{E}\|\delta_n^e\|^2 \\ &\leq \sum_{e=0}^{E-1} \sigma_n^2 \\ &= E\sigma_n^2, \end{aligned} \tag{24}$$

where we use Assumption 3 to obtain the upper bound in the second step. By Cauchy-Schwarz inequality, for the first term we have

$$\mathbb{E}\left\|\sum_{e=0}^{E-1}\nabla F_n(\psi_n^e)\right\|^2 \leq E\sum_{e=0}^{E-1}\mathbb{E}\|\nabla F_n(\psi_n^e)\|^2. \tag{25}$$

Under Assumption 4, we know that $\mathbb{E}\|\nabla F_n(\psi_n^e,\xi_n^e)\|^2 \leq G^2$. Since $\mathbb{E}\|\nabla F_n(\psi_n^e,\xi_n^e)\|^2 = \mathbb{E}\|\nabla F_n(\psi_n^e)+\delta_n^e\|^2 = \mathbb{E}\|\nabla F_n(\psi_n^e)\|^2+\mathbb{E}\|\delta_n^e\|^2$ and $\mathbb{E}\|\delta_n^e\|^2 \geq 0$, it follows that $\mathbb{E}[\|\nabla F_n(\psi_n^e)\|^2] \leq G^2$ and

$$\mathbb{E}\left\|\sum_{e=0}^{E-1}\nabla F_n(\psi_n^e)\right\|^2 \leq E\sum_{e=0}^{E-1}G^2 = E^2G^2. \tag{26}$$

Combing these two bounds with equation 23, we have

$$\begin{aligned}\mathbb{E}\|\psi_n^E-\psi_r\|^2 &\leq \eta_l^2(E^2G^2+E\sigma_n^2)\\ &= \eta_l^2E^2G^2+\eta_l^2E\sigma_n^2.\end{aligned} \tag{27}$$

This completes the proof. □

Lemma 2 (Deviation of Local Stochastic Gradients)**.** *Let Assumptions 1, 3 and 4 holds, the deviation between the average local stochastic gradient over E steps and the true gradient at the initial model ψ_r is bounded in expectation, i.e.,*

$$\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\nabla F_n(\psi_n^e,\xi_n^e)-\nabla F_n(\psi_r)\right\|^2 \leq \frac{2\sigma_n^2}{E}+\frac{2}{3}L^2\eta_l^2G^2E^2+L^2\eta_l^2\sigma_n^2E$$

Proof. We can decompose the total discrepancy into two terms, C_1 and C_2:

$$\begin{aligned}\frac{1}{E}\sum_{e=0}^{E-1}\nabla F_n(\psi_n^e,\xi_n^e)-\nabla F_n(\psi_r) = &\underbrace{\frac{1}{E}\sum_{e=0}^{E-1}(\nabla F_n(\psi_n^e,\xi_n^e)-\nabla F_n(\psi_n^e))}_{C_1}\\ &+\underbrace{\frac{1}{E}\sum_{e=0}^{E-1}(\nabla F_n(\psi_n^e)-\nabla F_n(\psi_r))}_{C_2}.\end{aligned} \tag{28}$$

Using the inequality $\|a+b\|^2 \leq 2\|a\|^2+2\|b\|^2$, we can bound the expected squared norm as

$$\mathbb{E}\,\|C_1+C_2\|^2 \leq 2\mathbb{E}\|C_1\|^2+2\mathbb{E}\|C_2\|^2. \tag{29}$$

We can rewrite the first term as

$$\mathbb{E}\|C_1\|^2 = \frac{1}{E^2}\mathbb{E}\left\|\sum_{e=0}^{E-1}(\nabla F_n(\psi_n^e,\xi_n^e)-\nabla F_n(\psi_n^e))\right\|^2 = \frac{1}{E^2}\mathbb{E}\left\|\sum_{e=0}^{E-1}\delta_n^e\right\|^2. \tag{30}$$

Using equation 24 again, and it follows that

$$\mathbb{E}\|C_1\|^2 \leq \frac{E\sigma_n^2}{E^2} = \frac{\sigma_n^2}{E}. \tag{31}$$

We can rewrite the second term as

$$\mathbb{E}\|C_2\|^2 = \mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}(\nabla F_n(\psi_n^e)-\nabla F_n(\psi_r))\right\|^2. \tag{32}$$

By Jensen's inequality, we have

$$\mathbb{E}\|C_2\|^2 \leq \frac{1}{E}\sum_{e=0}^{E-1}\mathbb{E}\|\nabla F_n(\psi_n^e) - \nabla F_n(\psi_r)\|^2$$
$$\leq \frac{L^2}{E}\sum_{e=0}^{E-1}\mathbb{E}\|\psi_n^e - \psi_r\|^2, \tag{33}$$

where we apply Assumption 1 to obtain the second inequality.

To bound the term $\mathbb{E}\left[\|\psi_n^e - \psi_r\|^2\right]$ for $e < E$, we use a result similar to Lemma 1, but for e steps instead of E, i.e.,

$$\mathbb{E}\|\psi_n^e - \psi_r\|^2 \leq \eta_l^2 e^2 G^2 + \eta_l^2 e\sigma_n^2. \tag{34}$$

Plugging this into Eq. equation 33, we have

$$\mathbb{E}\|C_2\|^2 \leq \frac{L^2}{E}\sum_{e=0}^{E-1}\left(\eta_l^2 e^2 G^2 + \eta_l^2 e\sigma_n^2\right) \tag{35}$$
$$= \frac{L^2\eta_l^2}{E}\left(G^2\sum_{e=0}^{E-1}e^2 + \sigma_n^2\sum_{e=0}^{E-1}e\right) \tag{36}$$

Using the formulas for the sum of integers and sum of squares, i.e., $\sum_{e=0}^{E-1} i = \frac{(E-1)E}{2}$ and $\sum_{e=0}^{E-1} i^2 = \frac{(E-1)E(2E-1)}{6}$, we have

$$\mathbb{E}\|C_2\|^2 \leq \frac{L^2\eta_l^2}{E}\left(G^2\frac{(E-1)E(2E-1)}{6} + \sigma_n^2\frac{(E-1)E}{2}\right)$$
$$\leq \frac{L^2\eta_l^2}{E}\left(G^2\frac{E^3}{3} + \sigma_n^2\frac{E^2}{2}\right)$$
$$= \frac{1}{3}L^2\eta_l^2G^2E^2 + \frac{1}{2}L^2\eta_l^2\sigma_n^2E. \tag{37}$$

Then we can combing equation 31 and equation 37 with equation 29 to obtain the bound for initial decomposition:

$$\mathbb{E}\left\|\frac{1}{E}\sum_{e=0}^{E-1}\nabla F_n(\psi_n^e,\xi_n^e) - \nabla F_n(\psi_r)\right\|^2 \leq 2\left(\frac{\sigma_n^2}{E}\right) + 2\left(\frac{1}{3}L^2\eta_l^2G^2E^2 + \frac{1}{2}L^2\eta_l^2\sigma_n^2E\right)$$
$$= \frac{2\sigma_n^2}{E} + \frac{2}{3}L^2\eta_l^2G^2E^2 + L^2\eta_l^2\sigma_n^2E. \tag{38}$$

This completes the proof. □

Lemma 3 (Expectation of Random Client Sampling)**.** *Let Assumption 5 holds and X_n be a client-specific random quantity independent of the client selection process, the expected expectation of the weighted sum over the randomly selected client set S_r satisfies that*

$$\mathbb{E}_{S_r}\left[\sum_{n\in S_r} w_n X_n\right] = \frac{M}{N}\sum_{n=1}^{N} w_n\mathbb{E}\left[X_n\right]$$

Proof. Let I_n be a binary random variable indicating the participation of client n in round r, where $I_n = 1$ if selected and 0 otherwise. With Assumption 5 holds, each client is selected independently with probability and it follows that $\mathbb{E}[I_n] = M/N$. The weighted sum over the randomly selected set S_r can be rewritten using these indicators such that

$$\sum_{n\in S_r} w_n X_n = \sum_{n=1}^{N} I_n w_n X_n. \tag{39}$$

Take the expectation of the above formula and applying linearity of expectation, we have

$$\mathbb{E}_{S_r}\left[\sum_{n\in S_r} w_n X_n\right] = \mathbb{E}\left[\sum_{n=1}^{N} I_n w_n X_n\right] = \sum_{n=1}^{N} \mathbb{E}\left[I_n w_n X_n\right]. \tag{40}$$

If the client selection I_n is independent of the client-specific quantity X_n, we have

$$\sum_{n=1}^{N} \mathbb{E}\left[I_n w_n X_n\right] = \sum_{n=1}^{N} \mathbb{E}\left[I_n\right] \cdot \mathbb{E}\left[w_n X_n\right] = \frac{M}{N} \sum_{n=1}^{N} w_n \mathbb{E}\left[X_n\right. \tag{41}$$

This completes the proof.

□

Theorem 2 (Global Convergence)**.** *Let Assumptions 1-6 hold and the server-side learning rate satisfies $\eta_\rho \leq 1/L$. After R communication rounds, the average expected squared norm of the global gradient is bounded as follows:*

$$\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\|\nabla F(\psi_r)\|^2 \leq \frac{F(\theta_0) - F^*}{R\Gamma_1} + \frac{\Gamma_2}{\Gamma_1},$$

where F^* is the minimum value of the global objective, Γ_1 and Γ_2 are constants that depend on problems' parameters and the algorithm's hyperparameters, but are independent of the total number of communication rounds R.

Proof. With Assumption 1 holds, we have

$$\mathbb{E}[F(\psi_{r+1})] \leq \mathbb{E}[F(\psi_r)] + \mathbb{E}\langle \nabla F(\psi_r), \psi_{r+1} - \psi_r\rangle + \frac{L}{2}\mathbb{E}\|\psi_{r+1} - \psi_r\|^2. \tag{42}$$

Due to our hybrid aggregation strategy, we decompose the update $\psi_{r+1} - \psi_r = (\theta_{r+1} - \theta_r, \rho_{r+1} - \rho_r)$ and the gradient $\nabla F(\psi_r) = (\nabla_\theta F(\psi_r), \nabla_\rho F(\psi_r))$, it follows that

$$\begin{aligned}
\mathbb{E}[F(\psi_{r+1})] &\leq \mathbb{E}[F(\psi_r)] + \mathbb{E}\langle \nabla_\theta F(\psi_r), \theta_{r+1} - \theta_r\rangle + \mathbb{E}\langle \nabla_\rho F(\psi_r), \rho_{r+1} - \rho_r\rangle \\
&\quad + \frac{L}{2}\mathbb{E}\|\theta_{r+1} - \theta_r\|^2 + \frac{L}{2}\mathbb{E}\|\rho_{r+1} - \rho_r\|^2 \\
&= \mathbb{E}[F(\psi_r)] + \underbrace{\left(\mathbb{E}[F(\theta_r, \rho_{r+1})] - F(\theta_r, \rho_r)\right)}_{D_1} \\
&\quad + \underbrace{\mathbb{E}\langle \nabla_\rho F(\psi_r), \rho_{r+1} - \rho_r\rangle + \frac{L}{2}\mathbb{E}\|\rho_{r+1} - \rho_r\|^2 - \left(\mathbb{E}[F(\theta_r, \rho_{r+1})] - F(\theta_r, \rho_r)\right)}_{D_2} \\
&\quad + \underbrace{\frac{L}{2}\mathbb{E}\|\theta_{r+1} - \theta_r\|^2}_{D_3} + \underbrace{\mathbb{E}\langle \nabla_\theta F(\psi_r), \theta_{r+1} - \theta_r\rangle}_{D_4}.
\end{aligned} \tag{43}$$

By Theorem 1, we know that

$$D_1 \leq -\frac{\eta_\rho}{2}\sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 + \frac{\eta_\rho U \zeta^2}{2}. \tag{44}$$

Under Assumption 1 and assuming θ_r fixed, we have

$$F(\theta_r, \rho_{r+1}) \leq F(\theta_r, \rho_r) + \langle \nabla_\rho F(\theta_r, \rho_r), \rho_{r+1} - \rho_r\rangle + \frac{L}{2}\|\rho_{r+1} - \rho_r\|^2. \tag{45}$$

Take the expectation of both sides and transpose the terms, we have

$$\mathbb{E}\langle \nabla_\rho F(\theta_r, \rho_r), \rho_{r+1} - \rho_r\rangle + \frac{L}{2}\mathbb{E}\|\rho_{r+1} - \rho_r\|^2 - \mathbb{E}[F(\theta_r, \rho_{r+1})] - F(\theta_r, \rho_r) \geq 0, \tag{46}$$

which is equivalent to $D_2 \geq 0$ and we can thus remove it from equation 43 to obtain the upper bound.

The update in D_3 is $\theta_{r+1} - \theta_r = \sum_{n \in S_r} w_n(\psi_n^E - \psi_r)$. Using Jensen's inequality, Lemma 1 and 3, we have

$$\begin{aligned} D_3 &= \frac{L}{2}\mathbb{E}\left\| \sum_{n \in S_r} w_n(\psi_n^E - \psi_r) \right\|^2 \leq \frac{LM}{2N} \sum_{n=1}^{N} w_n \mathbb{E}\|\psi_n^E - \psi_r\|^2 \\ &\leq \frac{LM}{2N} \sum_{n=1}^{N} w_n(\eta_l^2 G^2 E^2 + \eta_l^2 \sigma_n^2 E) \\ &= \frac{L^2 \eta_l^2 M}{2N}(G^2 E^2 + \bar{\sigma}^2 E), \end{aligned} \tag{47}$$

where $\bar{\sigma}^2 = \sum_{n=1}^{N} w_n \sigma_n^2$.

And we have $\theta_{r+1} - \theta_r = \sum_{n \in S_r} w_n(\psi_n^E - \psi_r) = -\eta_l \sum_{n \in S_r} w_n \sum_{e=0}^{E-1} \nabla_\theta F_n(\theta_n^e, \xi_n^e)$, using Assumption 3, we have

$$\begin{aligned} D_4 &= -\eta_l E \cdot \mathbb{E}\left\langle \nabla_\theta F(\psi_r), \sum_{n \in S_r} w_n \left(\frac{1}{E} \sum_{e=0}^{E-1} \nabla_\theta F_n(\psi_n^e, \xi_n^e) \right) \right\rangle \\ &= -\frac{\eta_l M}{N} E \cdot \left\langle \nabla_\theta F(\psi_r), \sum_{n=1}^{N} w_n \mathbb{E}\left[\frac{1}{E} \sum_{e=0}^{E-1} \nabla_\theta F_n(\psi_n^e, \xi_n^e) \right] \right\rangle \end{aligned} \tag{48}$$

Using the identity $-\langle a, b\rangle \leq \frac{1}{2}\|a-b\|^2 - \frac{1}{2}\|a\|^2$, we have

$$D_4 \leq -\frac{\eta_l EM}{2N}\|\nabla_\theta F(\psi_r)\|^2 + \frac{\eta_l EM}{2N}\mathbb{E}\left\| \nabla_\theta F(\psi_r) - \sum_{n=1}^{N} w_n \left(\frac{1}{E} \sum_{e=0}^{E-1} \nabla_\theta F_n(\psi_n^e, \xi_n^e) \right) \right\|^2. \tag{49}$$

Using Jensen's inequality and Lemma 2, we have

$$\begin{aligned} D_4 &\leq -\frac{\eta_l EM}{2N}\|\nabla_\theta F(\psi_r)\|^2 + \frac{\eta_l EM}{2N} \sum_{n=1}^{N} w_n \mathbb{E}\left\| \nabla_\theta F_n(\psi_r) - \left(\frac{1}{E} \sum_{e=0}^{E-1} \nabla_\theta F_n(\psi_n^e, \xi_n^e) \right) \right\|^2 \\ &\leq -\frac{\eta_l EM}{2N}\|\nabla_\theta F(\psi_r)\|^2 + \frac{\eta_l EM}{2N} \sum_{n=1}^{N} w_n \left(\frac{2\sigma_n^2}{E} + \frac{2}{3} L^2 \eta_l^2 G^2 E^2 + L^2 \eta_l^2 \sigma_n^2 E \right) \\ &= \frac{M}{N}\left(-\frac{\eta_l E}{2}\|\nabla_\theta F(\psi_r)\|^2 + \eta_l \bar{\sigma}^2 + \frac{1}{3}\eta_l^3 E^3 L^2 G^2 + \frac{1}{2}\eta_l^3 E^2 L^2 \bar{\sigma}^2 \right). \end{aligned} \tag{50}$$

Combing equation 44, equation 47 and equation 50 with equation 43, we have

$$\begin{aligned} \mathbb{E}[F(\psi_{r+1})] \leq \mathbb{E}[F(\psi_r)] &- \frac{\eta_l EM}{2N}\|\nabla_\theta F(\psi_r)\|^2 - \frac{\eta_\rho}{2} \sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 \\ &+ \frac{G^2 M}{N}\left(\frac{L^2 \eta_l^2 E^2}{2} + \frac{\eta_l^3 E^3 L^2}{3} \right) + \left(\eta_l + \frac{L^2 \eta_l^2 EM}{2N} + \frac{\eta_l^3 E^2 L^2 M}{2N} \right) \bar{\sigma}^2 \\ &+ \frac{\eta_\rho U \zeta^2}{2}. \end{aligned} \tag{51}$$

Note that

$$\sum_{u=0}^{U-1} \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_u)\|^2 \geq \mathbb{E}\|\nabla_\rho F(\theta_r, \rho_0)\|^2 = \|\nabla_\rho F(\theta_r, \rho_r)\|^2 = \|\nabla_\rho F(\psi_r)\|^2, \tag{52}$$

we can rewrite equation 51 as

$$\mathbb{E}[F(\psi_{r+1})] \leq \mathbb{E}[F(\psi_r)] - \frac{\eta_l EM}{2N}\|\nabla_\theta F(\psi_r)\|^2 - \frac{\eta_\rho}{2}\|\nabla_\rho F(\psi_r)\|^2 + \\ + \frac{G^2 M}{N}\left(\frac{L^2\eta_l^2 E^2}{2} + \frac{\eta_l^3 E^3 L^2}{3}\right) + \left(\eta_l + \frac{L^2\eta_l^2 EM}{2N} + \frac{\eta_l^3 E^2 L^2 M}{2N}\right)\bar{\sigma}^2 \\ + \frac{\eta_\rho U\zeta^2}{2}. \tag{53}$$

Let $\Gamma_1 = \min(\frac{\eta_l EM}{2N}, \frac{\eta_\rho}{2})$ and Γ_2 collects all constant terms, i.e., $\Gamma_2 = \frac{G^2 M}{N}\left(\frac{L^2\eta_l^2 E^2}{2} + \frac{\eta_l^3 E^3 L^2}{3}\right) + \left(\eta_l + \frac{L^2\eta_l^2 EM}{2N} + \frac{\eta_l^3 E^2 L^2 M}{2N}\right)\bar{\sigma}^2 + \frac{\eta_\rho U\zeta^2}{2}$, we can rearrange the inequality as

$$\Gamma_1 \mathbb{E}\|\nabla F(\psi_r)\|^2 \leq \mathbb{E}[F(\psi_r)] - \mathbb{E}[F(\psi_{r+1})] + \Gamma_2. \tag{54}$$

Summing over $r = 0, \ldots, R-1$, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Gamma_1\mathbb{E}\|\nabla F(\psi_r)\|^2 &\leq \sum_{r=0}^{R-1}(\mathbb{E}[F(\psi_r)] - \mathbb{E}[F(\psi_{r+1})]) + \sum_{r=0}^{R-1}\Gamma_2 \\ &= F(\psi_0) - \mathbb{E}[F(\theta_R)] + R\,\Gamma_2 \\ &\leq F(\psi_0) - F^* + R\,\Gamma_2,\end{aligned} \tag{55}$$

where F^* is the minimum value of the global objective. Dividing by $R\,\Gamma_1$ gives the final result

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla F(\psi_r)\|^2 \leq \frac{F(\psi_0) - F^*}{R\,\Gamma_1} + \frac{\Gamma_2}{\Gamma_1}. \tag{56}$$

This completes the proof.

□

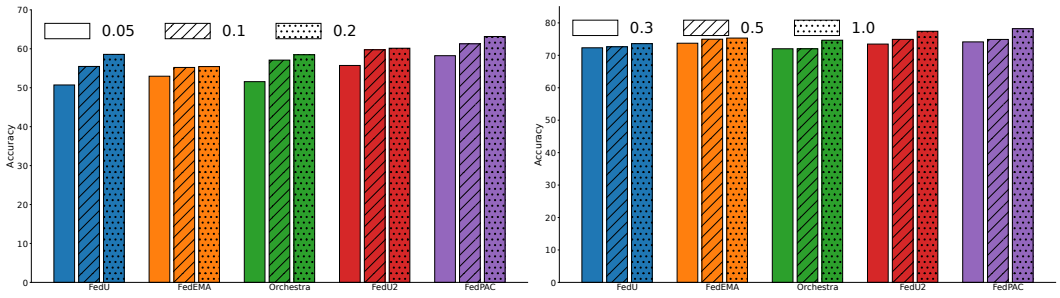
C Additional Results

C.1 Sensitivity to hyper-parameters

We now continue that analysis of sensitivity to key federated learning hyper-parameters, most of the experiments are conducted under cross-device setting on CIFAR-10 and CIFAR-100. **First**, regarding the client participation rate, as shown in Figure 4, performance in the cross-silo setting remained stable with only a marginal drop at lower rates. In the cross-device setting, while all methods experience performance degradation with lower participation rates, FedPAC exhibited the highest resilience. **Furthermore**, we assessed the impact of the total number of clients and plot the result in Figure 5. It can be viewed that FedU2 achieves highest accuracy in settings with less clients, while FedPAC demonstrates superior robustness, maintaining more stable performance as the total number of clients changes. **Finally**, we analyze the effect of local training epochs in Figure 6. A reduction in local epochs led to significant performance degradation for all methods. In contrast, FedPAC maintained high accuracy across different epoch settings and could achieve a comparable level of accuracy to competing methods but with a reduced number of local epochs. **Collectively**, these experiments demonstrate he robustness of FedPAC to variations in the training configuration.

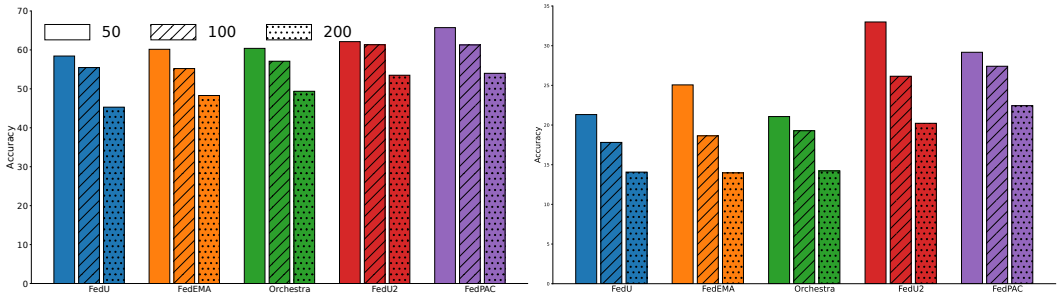
C.2 Representation Visualization

To provide an intuitive and qualitative assessment of the learned representations, we visualize the feature embeddings using t-SNE. We first examine the problem of semantic misalignment by comparing local and global models from both FedU2 and FedPAC. As shown in Figure 7, while local models in FedU2 learn locally coherent representations, their feature spaces are misaligned with one another. Consequently, in the aggregated global model's feature space, representations from different categories may become mixed and thus reduce discriminability. In contrast, FedPAC learns



(a) Cross-Device (b) Cross-Silo

Figure 4: Sensitivity to participation ratio on cross-device (left) and cross-silo (right) settings. FedPAC maintains stable accuracy even with low participation rates.



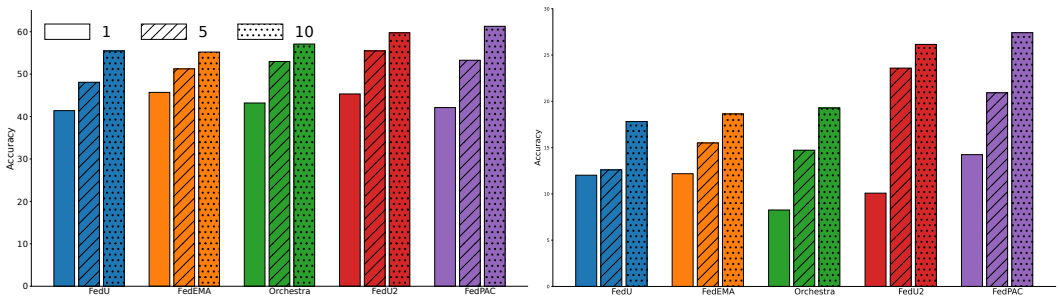
(a) CIFAR-10 (b) CIFAR-100

Figure 5: Sensitivity to clients number on CIFAR-10 (left) and CIFAR-100 (right) settings.

maintains semantic consistency across clients, forming a global representation space with enhanced inter-class separation, demonstrating its ability to mitigate representation drift. Then we compare the final global representations learned by all methods in Figure 8. This visualization shows that FedPAC learns representations with better intra-class compactness and inter-class separability, providing further evidence of the high discriminative power of the feature space cultivated by our framework.

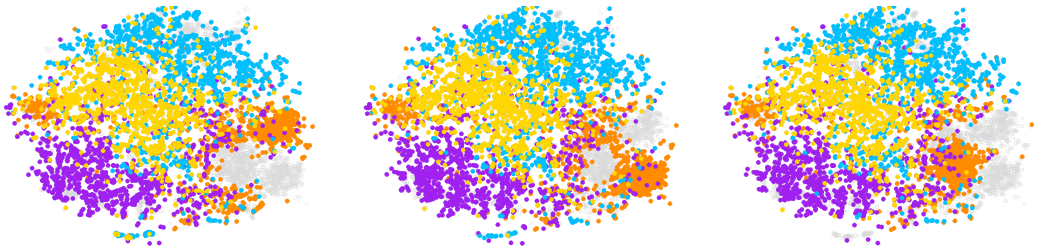
D THE USE OF LARGE LANGUAGE MODELS (LLMS)

During the writing process, we used LLMs to identify grammatical errors in the article and polish some sentences.

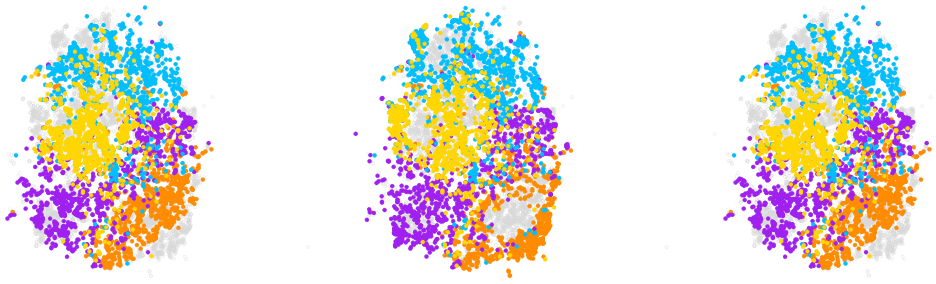


(a) CIFAR-10 (b) CIFAR-100

Figure 6: Sensitivity to local epochs on CIFAR-10 (left) and CIFAR-100 (right). FedPAC shows consistent robustness and efficiency across all settings.

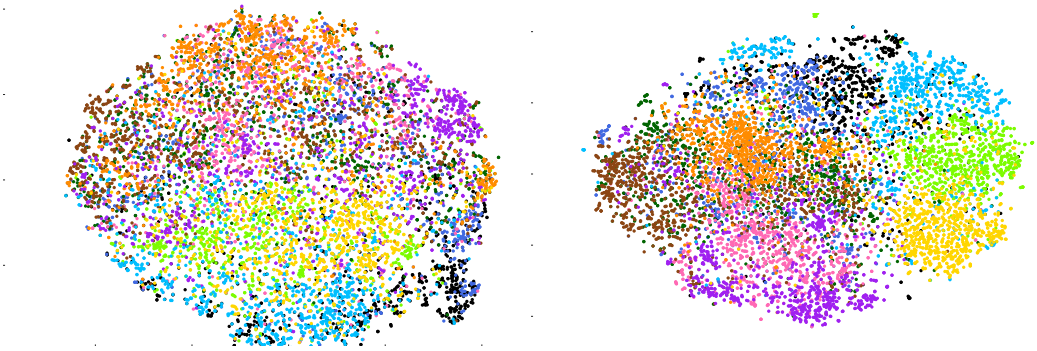


(a) FedU2-local 1 (b) FedU2-local 2 (c) FedU2-global



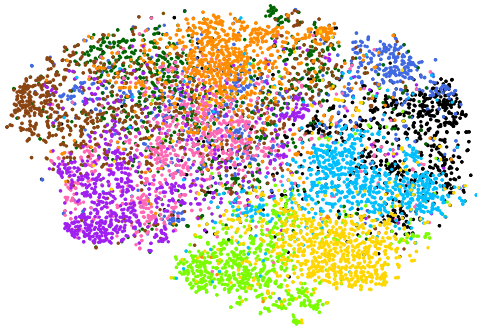
(d) FedPAC-local 1 (e) FedPAC-local 2 (f) FedPAC-global

Figure 7: Visualization of local and global representations, demonstrating that FedPAC alleviates cross-client representation drift, leading to enhanced global consistency compared to the FedU2.

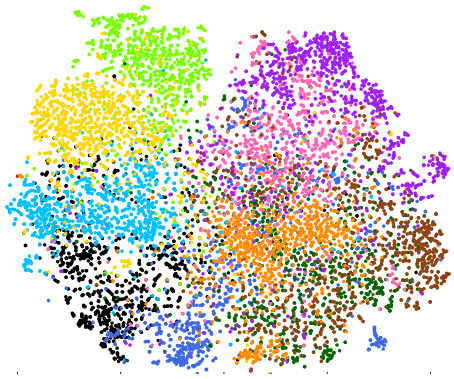


(a) FedEMA

(b) FedU2-local 2



(c) FedU2



(d) FedPAC

Figure 8: Visualization of the representation space learned by the global model of different methods.