Harmonizing Federated Heterogeneous Optimization via Adaptive Objective Rectification

Jianrong Lu[1] , **Bangwei Li**[1] , **Zhuoya Gu**[1] , **Peng Fang**[1] ,
Ziming Zhao[1] and **Jianhai Chen**[1,*]
[1]Zhejiang University
jrong.alvin@gmail.com, libangwei@zju.edu.cn, allenya3@gmail.com,
{fangpeng, zhaoziming, chenjh919}@zju.edu.cn

Abstract

Federated optimization under data heterogeneity presents a significant challenge, often leading to suboptimal model performance. While numerous methods aim to replicate the ideal performance of centralized training, they frequently fall short in highly heterogeneous settings. In this paper, we introduce HaFedHo, an adaptive objective rectification method that harmonizes local training with the ideal data-centralized objective, requiring minimal modifications to the standard federated learning framework. HaFedHo operates by first decoupling the centralized objective and then employing a dynamic Taylor series expansion to accurately estimate the global objective for each client. Our theoretical analysis shows that the estimation error provably converges to zero as training progresses. Furthermore, extensive experiments on real-world datasets demonstrate that HaFedHo surpasses state-of-the-art methods, including SCAFFOLD, MimeLite, and FedDyn, in both test accuracy and communication efficiency. Notably, HaFedHo maintains its superior performance even with a client participation rate as low as 0.2% in severely heterogeneous environments.

1 Introduction

Federated Learning (FL) [McMahan *et al.*, 2017] has emerged as a dominant distributed learning blueprint, empowering a centralized coordinator to orchestrate global model optimization across edge clients without centralizing their raw assets. By anchoring training data securely on-device, FL fundamentally satisfies modern strict privacy mandates. This architecture has matured beyond theoretical exercises to power large-scale commercial deployments by technology pioneers such as Google and Apple, while demonstrating immense promise in privacy-critical sectors including finance, digital healthcare [Lu *et al.*, 2026], and smart cities.

Despite its structural elegance, executing FL in realistic ecosystems exposes a fundamental vulnerability: the *Information Asymmetry Paradox*. Because individual clients experience distinct contexts, their local data repositories are inherently non-Independent and Identically Distributed (non-IID) in terms of distribution and volume. Compounding this is the necessity of partial client participation, where only an unpredictable fraction of nodes remains active during any single round. Consequently, local optimization vectors drift along divergent trajectories, causing severe federated heterogeneous optimization anomalies that degrade global model performance [Sun *et al.*, 2025; Li *et al.*, 2020b] and lead to outcomes far inferior to traditional centralized training [Wang *et al.*, 2020b; Acar *et al.*, 2021; Lu *et al.*, 2025].

To bridge this deficit, contemporary solutions typically attempt to constrain local updates to prevent them from straying from the global pathway [Bakhtiari *et al.*, 2020; Yoon *et al.*, 2021]. However, enforcing rigid behavioral control within volatile environments forces state-of-the-art frameworks to depend on brittle, highly idealized assumptions. For instance, landmark protocols like SCAFFOLD [Karimireddy *et al.*, 2020] and FedDyn [Acar *et al.*, 2021] require stable client participation to properly calibrate their internal control variates or dynamic regularizers. Similarly, Mime [Karimireddy *et al.*, 2021] relies on relatively homogeneous gradient spaces, whereas FedSplit [Pathak and Wainwright, 2020] collapses outside the narrow confines of full client engagement and strictly convex objectives. When confronted with real-world, deeply fragmented heterogeneous data, the efficacy of these methods deteriorates rapidly.

Departing from conventional reactive drift correction, we propose HaFedHo, a proactive paradigm designed to natively align decentralized execution with centralized optimization. The core innovation of HaFedHo is the explicit rectification of each client's local objective function *before* training begins. We achieve this by compelling each client to optimize an adaptive, Taylor-approximated surrogate of the true, data-centralized global objective. By isolating the local function and projecting the missing collaborative topologies via an agile, first-order Taylor series expansion (refreshed dynamically via cross-round remnants), HaFedHo effectively removes the information barrier between silos.

Backed by rigorous theoretical guarantees, we prove the convergence properties of HaFedHo across both convex and complex non-convex surfaces under extreme data non-IID conditions. Empirical validations across three standard benchmarks (FEMNIST, CIFAR-10, and CIFAR-100) con-

*Corresponding author.

firm that HaFedHo consistently outperforms existing strong baselines such as SCAFFOLD [Karimireddy *et al.*, 2020], MimeLite [Karimireddy *et al.*, 2021], MOON [Li *et al.*, 2021], and FedDyn [Acar *et al.*, 2021], showcasing remarkable resilience in scenarios defined by severe data sparsity and ultra-low node availability.

Our core contributions are summarized as follows:

- We present a novel optimization framework that transforms isolated local training by embedding a Taylor-series approximation of the missing global data landscape directly into the local objective function, supported by comprehensive convergence proofs.
- We introduce a practical, communication-efficient implementation of this framework that dynamically updates its approximation of the global objective, and mathematically demonstrate that its internal estimation error decays toward zero as the iterations progress.
- Through extensive empirical evaluation, we demonstrate that HaFedHo yields superior accuracy trajectories and accelerated convergence behavior compared to state-of-the-art optimizers, particularly in highly adverse client participation environments.

2 Preliminaries

2.1 Federated Learning

We study a standard decentralized ecosystem comprising a focal server and n isolated edge nodes, where each node $i \in [n]$ restricts computation to its private domain D_i. The objective is to cooperatively resolve a global empirical risk minimization problem under a uniform consensus:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\Gamma(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{w}) \right], \tag{1}$$

where $f_i(\boldsymbol{w}) = \frac{1}{|D_i|}\sum_{\pi\in D_i} \ell(\boldsymbol{w}, \pi)$ characterizes the localized loss surface over the node's distinct data partition.

In any active communication round t, participants pull the consensus state to initialize their local models ($\boldsymbol{w}_i^{(t,0)} = \boldsymbol{w}^t$) and advance across λ sequential local epochs using Stochastic Gradient Descent (SGD):

$$\boldsymbol{w}_i^{(t,\lambda)} = \boldsymbol{w}_i^{(t,0)} - \eta \boldsymbol{g}_i^t, \quad \text{with} \quad \boldsymbol{g}_i^t = \sum_{k=0}^{\lambda-1} \nabla f_i(\boldsymbol{w}_i^{(t,k)}). \tag{2}$$

Upon completing local trajectories, the server orchestrates a global synchronization step matching the FedAvg protocol [McMahan *et al.*, 2017]:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \frac{\eta}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^t. \tag{3}$$

2.2 Data-centralized Training

Conversely, the unconstrained data-centralized baseline operates without information silos. The master server directly operates on the joint pool $D = \bigcup_{i=1}^{n} D_i$ by executing optimization directly over the global data distribution:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \left[\Xi(\boldsymbol{w}, D) = \frac{1}{|D|}\sum_{\pi\in D} \ell(\boldsymbol{w}, \pi) \right]. \tag{4}$$

Centralized optimization shifts the parameters iteratively through identical sequential update steps:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \boldsymbol{d}^t, \quad \text{with} \quad \boldsymbol{d}^t = \sum_{k=0}^{\lambda-1} \nabla \Xi(\boldsymbol{w}^{(t,k)}), \tag{5}$$

where $\boldsymbol{d}^t$ tracks the true global gradient trajectory over the fully unified loss landscape Ξ.

3 Related Works

We group them by how they handle the global information gap across isolated client nodes.

Communication-Intensive Variance Reduction. The most direct way to eliminate localized client drift is to restrict the local optimization horizon to a single step, as seen in protocols like FedSMB [Bakhtiari *et al.*, 2020] and FedSGD [McMahan *et al.*, 2017]. This ensures that the global update path exactly matches centralized training. However, this strategy introduces massive communication overhead and renders the training process highly unstable when paired with modern deep architectures that rely on stateful layers like batch normalization or dropout. Consequently, modern FL research focuses on extending the local training duration while correcting the resulting client drift.

History-Driven Trajectory Realignment. One prominent approach uses historical optimization metadata to adjust local updates and smooth out cross-client variances [Hsu *et al.*, 2019; Karimireddy *et al.*, 2021; Khanduri *et al.*, 2021; Xu *et al.*, 2021; Karimireddy *et al.*, 2020; Chen *et al.*, 2024]. For example, SCAFFOLD [Karimireddy *et al.*, 2020] introduces stateful control variates to track client drift, while Mime and its lightweight counterpart MimeLite [Karimireddy *et al.*, 2021] replace local momentum terms (originally explored in FedavgM [Hsu *et al.*, 2019]) with precise global gradient vectors collected across all participants. While effective on uniform datasets like FEMNIST [Caldas *et al.*, 2018], these methods struggle in highly volatile environments with sporadic client availability. Furthermore, transmitting historical state vectors increases communication costs and heightens the risk of information leakage, making systems more vulnerable to advanced gradient inversion attacks [Lu *et al.*, 2024; Geiping *et al.*, 2020; Lu *et al.*, 2023].

State-Constrained Optimization Anchoring. Another strategy adds regularizers, proximal terms, or feature alignment structures directly to the local loss functions to penalize deviations from the global archetype [Li *et al.*, 2020a; Gao *et al.*, 2022; Acar *et al.*, 2021; Wang *et al.*, 2024; Hu *et al.*, 2024; Chu *et al.*, 2025]. For instance, FedProx [Li *et al.*, 2020a] introduces a proximal term to bound local drift, while FedDyn [Acar *et al.*, 2021] uses dynamic quadratic penalty terms to align local minima with the global objective under stable participation assumptions. Similarly, FedSplit [Pathak and Wainwright, 2020] applies operator splitting methods to tackle non-IID data distributions, and contemporary objective-driven approaches [Gao *et al.*, 2022; Wang *et al.*, 2024; Hu *et al.*, 2024; Chu *et al.*, 2025] regularize local spaces via features or constraints. However, these methods act as reactive patches; they try to control client drift *after* the client objective functions have already diverged from the

centralized target. In contrast, our approach directly approximates the global data-centralized objective from the start, ensuring that local optimization vectors naturally align with the centralized target.

Post-Hoc Topology Reshaping. A final category of work focuses entirely on the server side, modifying the aggregation phase to minimize the impact of non-IID data distributions and domain shifts [Wang *et al.*, 2020a; Chen and Chao, 2020; Xu *et al.*, 2024; Ning *et al.*, 2024; Zeng *et al.*, 2024; Shen *et al.*, 2025; Zheng *et al.*, 2025]. Notable examples include FedBN [Li *et al.*, 2020b], which excludes batch normalization layers from the aggregation step, and FedLWS [Shi *et al.*, 2025a; Shi *et al.*, 2025b], which introduces adaptive weight rebalancing based on client features. Alternative aggregation paradigms further reshape the global landscape via Bayesian ensembling [Chen and Chao, 2020], layer matching [Wang *et al.*, 2020a], or pluggable client scheduling filters [Xu *et al.*, 2024; Ning *et al.*, 2024; Zeng *et al.*, 2024; Shen *et al.*, 2025; Zheng *et al.*, 2025]. Because our proposed method operates entirely during the local training phase, it is complementary to these server-side techniques, variance-reducing momentum methods [Wang *et al.*, 2023], Sharpness-Aware Minimization (SAM) variants [Sun *et al.*, 2023; Xing *et al.*, 2025].

4 Methodology

4.1 Motivation

Our approach is motivated by a simple principle: if each client's local optimization task closely mirrors the ideal centralized training objective, the overall FL process should yield a better global model.

We observe that if we expand the data-centralized objective $\Xi(\boldsymbol{w})$ in a Taylor series around the place where $\boldsymbol{w} = \boldsymbol{w}^t$ (i.e., $\Xi(\boldsymbol{w}) = \Xi(\boldsymbol{w}^t) + [\nabla\Xi(\boldsymbol{w}^t)]^\top(\boldsymbol{w} - \boldsymbol{w}^t) + \mathcal{O}(\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2)$, where $\mathcal{O}(\|\boldsymbol{w} - \boldsymbol{w}^t\|_2^2)$ is the Peano's remainder error), the Peano's remainder error would gradually vanish as the iteration increases. This suggests that if each client ignores the Peano's remainder error and solves the first-order Taylor's series approximation, all the clients will converge to the stationary point of $\Xi(\boldsymbol{w})$. Therefore, we have the following ideal way to imitate the data-centralized training.

Proposition 1. *Suppose all clients have access to the global loss $\Xi(\boldsymbol{w}^t)$ and the full-batch centralized gradient $\nabla\Xi(\boldsymbol{w}^t)$. If each client i minimizes the local objective*

$$\varphi_i(\boldsymbol{w}) = \Xi(\boldsymbol{w}^t) + \nabla\Xi(\boldsymbol{w}^t)^\top(\boldsymbol{w} - \boldsymbol{w}^t), \tag{6}$$

then the federated optimization process converges to a stationary point of $\Xi(\boldsymbol{w})$.

Proof. See supplementary material. □

While Proposition 1 provides a theoretical ideal, it is not directly applicable in FL because clients cannot compute the global quantities $\Xi(\boldsymbol{w}^t)$ and $\nabla\Xi(\boldsymbol{w}^t)$. In the following section, we present a practical method to approximate this ideal objective.

Algorithm 1: A Complete Description of HaFedHo

Input : T; $\boldsymbol{w}_i^{(0,0)} = \boldsymbol{w}^0$; $\boldsymbol{g}_i^0 = \boldsymbol{g}^0$; $\hat{f}_i^0 = \hat{f}^0 = 0$; learning rate η; the number of SGD steps λ; penalty factor μ.
Output : The global model $\boldsymbol{w}^T$.
for $t = 1, 2, \cdots, T$ **do**
  Randomly select m clients $\mathcal{M} \subseteq \{1, 2, \cdots, n\}$.
  The server broadcasts $\boldsymbol{g}^{t-1}$, $\hat{f}^{t-1}$ to clients.
  for ***client*** $i \in \mathcal{M}$ **do**
    $\boldsymbol{w}^t = \boldsymbol{w}_i^{(t,0)} = \boldsymbol{w}_i^{(t-1,0)} - \eta\boldsymbol{g}^{t-1}$.
    if *client i is sampled in the last round* **then**
      $\boldsymbol{h} = \boldsymbol{g}^{t-1} - p_i\boldsymbol{g}_i^{t-1}$.
      $\hat{f} = \hat{f}^{t-1} - p_i\hat{f}_i^{t-1}$.
    else
      $\boldsymbol{h} = \boldsymbol{g}^{t-1}$.
      $\hat{f} = \hat{f}^{t-1}$.
    Define local objective:
    $\varphi_i(\boldsymbol{w}) = f_i(\boldsymbol{w}) + \mu\Big(\hat{f} + \boldsymbol{h}^\top(\boldsymbol{w} - \boldsymbol{w}^t)\Big)$.
    Perform λ local SGD steps:
    $\boldsymbol{w}_i^{(t,k+1)} \leftarrow \boldsymbol{w}_i^{(t,k)} - \eta\nabla\varphi_i(\boldsymbol{w}_i^{(t,k)})$.
    Compute average local loss:
    $\hat{f}_i^t \leftarrow \frac{1}{\lambda}\sum_{k=0}^{\lambda-1} f_i(\boldsymbol{w}_i^{(t,k)})$.
    Compute total local gradient:
    $\boldsymbol{g}_i^t \leftarrow \sum_{k=0}^{\lambda-1}\nabla\varphi_i(\boldsymbol{w}_i^{(t,k)})$.
    Client i uploads $\hat{f}_i^t$ and $\boldsymbol{g}_i^t$ to the server.
  $\hat{f}^t = \sum_{i=1}^m \frac{1}{m}\hat{f}_i^t$.
  $\boldsymbol{g}^t = \sum_{i=1}^m \frac{1}{m}\boldsymbol{g}_i^t$.

4.2 HaFedHo: The Proposed Method

The optimal guarantee in Proposition 1 inspires us to develop a method from the local objective perspective. However, since clients lack knowledge of the entire dataset D, it is impractical for each client to employ the Taylor estimation in Eq. (6) and calculate the full-batch gradient $\nabla\Xi(\boldsymbol{w}^t)$. To overcome this challenge, our core idea is to decouple $\Xi(\boldsymbol{w}, D)$ into multiple client-level objective functions, i.e., $f_i(\boldsymbol{w}, D_i), i \in \{1, 2, \cdots, n\}$, according to the datasets of clients and then dynamically approximate each $f_i(\boldsymbol{w}, D_i)$ with the corresponding gradients.

Global-Local Objective Decomposition. To bridge the gap between federated execution and centralized training, we must first isolate the local optimization landscape from the overarching global target. The empirical risk of the centralized model $\Xi(\boldsymbol{w}, D)$ can be inherently decomposed into the active client's local loss and the aggregate loss of all unseen peers:

$$\Xi(\boldsymbol{w}, D) = p_i f_i(\boldsymbol{w}, D_i) + \sum_{j\neq i}^{n} p_j f_j(\boldsymbol{w}, D_j), \tag{7}$$

where $p_i = \frac{|D_i|}{|D|}$ represents the data fraction. Directly optimizing this joint function is impossible due to strict data silos. While one could theoretically approximate the unseen term $\sum_{j\neq i}^n p_j f_j(\boldsymbol{w}, D_j)$ using a first-order Taylor expansion around the current global weights $\boldsymbol{w}^t$, requesting

real-time gradients $\nabla f_j(\boldsymbol{w}^t)$ across the entire network is both communication-prohibitive and a severe privacy risk.

First-Order Proxy via Historical Trajectories. Rather than forcing costly synchronizations, we can construct a surrogate for the missing gradient information using historical trajectory data. During standard model aggregation, the server already compiles the global update vector $\boldsymbol{g}^{t-1} = \sum_{i=1}^{n} p_i \boldsymbol{g}_i^{t-1}$. This aggregated vector serves as a natural, albeit slightly delayed, proxy for the true centralized gradient.

By projecting the non-local components using this historical proxy, client i constructs a locally computable surrogate objective. To isolate the exact contribution of the *other* clients and prevent self-reinforcement, we subtract the active client's previously uploaded gradient $\boldsymbol{g}_i^{t-1}$ from the global average:

$$\varphi_i(\boldsymbol{w}) = p_i f_i(\boldsymbol{w}, D_i) + \sum_{j \neq i}^{n} p_j f_j(\boldsymbol{w}^t) + [\boldsymbol{g}^{t-1} - p_i \boldsymbol{g}_i^{t-1}]^\top (\boldsymbol{w} - \boldsymbol{w}^t). \tag{8}$$

Applying identical arithmetic to the scalar loss values—estimating the unseen global loss $\sum_{j \neq i}^{n} p_j f_j(\boldsymbol{w}^t)$ via $\sum_{i=1}^{n} p_i f_i(\boldsymbol{w}^{t-1}) - p_i f_i(\boldsymbol{w}^{t-1})$—transforms the objective into a fully computable state:

$$\begin{aligned} \varphi_i(\boldsymbol{w}) = & p_i f_i(\boldsymbol{w}, D_i) + [f^{t-1} - p_i f_i^{t-1}] \\ & + [\mathbf{g}^{t-1} - p_i \mathbf{g}_i^{t-1}]^\top (\boldsymbol{w} - \boldsymbol{w}^t), \end{aligned} \tag{9}$$

where $f^{t-1} = \sum_{i=1}^{n} p_i f_i^{t-1}$ and $f_i^{t-1} = f_i(\boldsymbol{w}^{t-1})$. Because f_i is a zero-order scalar, transmitting it to the server imposes near-zero bandwidth overhead. Furthermore, protecting these scalars via cryptographic protocols (such as Secure Multi-Party Computation) is computationally trivial compared to encrypting high-dimensional gradient matrices, ensuring robust privacy preservation.

Mitigating Stale States in Partial Participation. Operating under realistic federated regimes dictates that only a small fraction of the client pool engages per round. Consequently, the global aggregates $\hat{f}^{t-1}$ and $\boldsymbol{g}^{t-1}$ are inherently biased by the sampled cohort.

A critical overlap issue arises if client i is sampled in successive rounds: its local state is already baked into the global broadcast, risking an over-correction or momentum bias. To neutralize this, we selectively subtract the local historical terms ($p_i \boldsymbol{g}_i^{t-1}$ and $p_i \hat{f}_i^{t-1}$) *only* if the client participated in round $t-1$. For newly sampled clients, the raw global aggregates are utilized directly, as their previous updates are not present in the immediate global state.

Finally, to stabilize training dynamics and scale the impact of the baseline local loss, we introduce a tuning parameter μ to modulate the intensity of the proxy terms. This yields the ultimate minimization target for the federated system:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \left[\Gamma(\boldsymbol{w}) := \sum_{i=1}^{n} p_i \varphi_i(\boldsymbol{w}) \right], \tag{10}$$

where the client-side surrogate evaluates to $\varphi_i(\boldsymbol{w}) = f_i(\boldsymbol{w}, D_i) + \mu \left[\hat{f}^{t-1} \right] + \mu \left[\boldsymbol{h} \right]^\top (\boldsymbol{w} - \boldsymbol{w}^t)$. The precise mechanics for determining the dynamically adjusted scalars $\hat{f}$ and vectors $\boldsymbol{h}$ based on participation history are detailed in Algorithm 1, while Figure 1 illustrates the geometric intuition guiding this optimization trajectory.

4.3 Theoretical Analysis

We analyze the convergence properties of HaFedHo under standard assumptions from the FL literature [Li *et al.*, 2020a; Karimireddy *et al.*, 2020; Karimireddy *et al.*, 2021; Gao *et al.*, 2022; Acar *et al.*, 2021].

Assumption 1 (Bounded Dissimilarity)**.** *For a positive constant $B \geq 1$, we assume that the variance of local gradients with respect to the global gradient is bounded: $\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{w})\|^2 \leq B^2 \|\nabla \Gamma(\boldsymbol{w})\|^2, \forall \boldsymbol{w}$.*

Assumption 2 (Bounded Variance)**.** *The stochastic gradient $\nabla f_i(\boldsymbol{w}; \zeta_i)$ computed on a mini-batch ζ_i is an unbiased estimator of the true local gradient $\nabla f_i(\boldsymbol{w})$, with bounded variance: $\mathbb{E}_{\zeta_i} \left[\|\nabla f_i(\boldsymbol{w}; \zeta_i) - \nabla f_i(\boldsymbol{w})\|^2 \right] \leq \sigma^2, \forall i, \boldsymbol{w}$.*

Assumption 3 (Lipschitz Smoothness)**.** *Each local loss function f_i is β-smooth, meaning its gradient is Lipschitz continuous: $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\| \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|, \forall i, \boldsymbol{x}, \boldsymbol{y}$.*

Definition 1 (Estimation Error)**.** *We use local gradient $\boldsymbol{g}_i^t$ to estimate $\nabla f_i(\boldsymbol{w}^t)$. Thus, the quantity $\boldsymbol{h}$ defined as:*

$$\boldsymbol{h} = \begin{cases} \boldsymbol{g}^t - p_i \boldsymbol{g}_i^t, & \text{if client } i \text{ is sampled,} \\ \boldsymbol{g}^t, & \text{otherwise,} \end{cases} \tag{11}$$

actually estimates $\nabla \Gamma(\boldsymbol{w}^t)$. Besides, we utilize a hyperparameter μ to control the magnitude of $\boldsymbol{h}$, therefore, the estimation error Θ is formally defined as:

$$\Theta := \mathbb{E}_t \left[\left\| \mu \boldsymbol{h} - \nabla \Gamma(\boldsymbol{w}^t) \right\|_2^2 \right] \tag{12}$$

Theorem 1 (Estimation Error Bound)**.** *Under Assumptions 1-3, for a non-convex objective, client sampling ratio $\frac{m}{n}$, T communication rounds, λ local steps, and a sufficiently small learning rate $\eta = \mathcal{O}(\frac{1}{T})$, the estimation error Θ is bounded:*

$$\begin{aligned} \Theta \leq & C_1 \mathcal{O}\left(\frac{\sigma^4}{m \lambda T^{\frac{5}{2}}} + \frac{B^2}{T^3} \right) + C_2 \mathcal{O}\left(\frac{\sigma^4}{m \lambda \sqrt{T}} + \frac{B^2}{T} \right) \\ & + C_3 \mathcal{O}\left(\frac{1}{T^2} \right) + C_4 \mu^2, \end{aligned} \tag{13}$$

where $C_1 = 18\mu^2\beta^2 B^2 (1 + \frac{m}{n} p_i^2 \lambda^2)$, $C_2 = (\mu - 1)^2 + \frac{m}{n}(\lambda^2 \mu^2 p_i^2 B^2 + 2)$, $C_3 = 9\lambda\mu^2\beta^2\sigma^2(1 + \frac{m}{n} p_i^2)$ and $C_4 = \sigma^2(1 + \frac{m}{n} p_i^2 \lambda^2)$.

Proof. See supplementary material. □

Analysis: Theorem 1 characterizes the effectiveness of using the latest global gradients $\boldsymbol{g}^t$ and the local gradient $\boldsymbol{g}_i^t$ to estimate all the local gradients $\sum_{j \neq i}^{n} \nabla f_j(\boldsymbol{w}^t)$ of other clients. As the number of rounds T approaches infinity, the first three terms in the upper bound of the error gradually vanish, leaving only the last term $C_4 \mu^2$. Since μ can be adjusted as a hyperparameter inversely proportional to the iteration round t, $C_4 \mu^2$ also tends to zero as the iterations progress. This indicates that as HaFedHo converges (Theorem 2), the estimation error becomes extremely small, almost negligible.

Theorem 2 (Convergence)**.** *Consider HaFedHo running for T rounds, with λ local SGD steps, and a client sampling ratio of $\frac{m}{n}$. If the learning rate η is sufficiently small, then the average squared gradient norm is bounded as follows:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla\Gamma(\boldsymbol{w}^t)\|^2\right] \leq \frac{\Gamma(\boldsymbol{w}^0)-\Gamma^*}{T\cdot C_G}+\frac{C_\Sigma}{C_G}$$

where $\Gamma(\boldsymbol{w}^0)-\Gamma^*$ is the initial sub-optimality, and the constants C_G and C_Σ are constants (see the supplementary).

Proof. See supplementary material. □

5 Experiments

5.1 Experimental Setup

We implement our evaluation framework using FedML [He *et al.*,].

Dirichlet Data Partitioning. Following the standard FedML protocol, we simulate statistical heterogeneity across n clients via Latent Dirichlet Allocation. Specifically, we sample $\boldsymbol{p}_l \sim \text{Dirichlet}_n(q)$, assigning a fraction $\boldsymbol{p}_{l,i}$ of class l instances to client i. The concentration parameter q dictates the degree of non-IIDness, with lower values yielding higher statistical heterogeneity.

Baselines. Our approach is benchmarked against eight state-of-the-art federated learning algorithms: MOON [Li *et al.*, 2021], FedProx [Li *et al.*, 2020a], pFedAFM [Yi *et al.*, 2025], FedDyn [Acar *et al.*, 2021], MimeLite [Karimireddy *et al.*, 2021], FedLWS [Shi *et al.*, 2025a], SCAFFOLD [Karimireddy *et al.*, 2020], and FedAvg [McMahan *et al.*, 2017].

Hyperparameters. For the FEMNIST dataset, we configure a total of $n = 3400$ clients with a participation rate of 10%. Conversely, evaluations on CIFAR-10 and CIFAR-100 utilize $n = 100$ clients, maintaining the 10% sampling fraction. *Optimal hyperparameters for all baseline methods were rigorously tuned for fair comparison (refer to the supplementary material).*

Evaluation Metric. Performance is quantified using *test accuracy*, defined as the proportion of correctly predicted test instances by the aggregated global model.

5.2 Performance Evaluation

Test Accuracy. Table 1 benchmarks the predictive accuracy across three non-IID datasets. HaFedHo establishes a clear state-of-the-art, consistently surpassing all baselines. Standard methods (e.g., FedAvg, FedProx, SCAFFOLD, MOON) plateau at similar levels. Notably, while FedDyn achieves competitive peaks, its performance is highly erratic, dropping below standard FedAvg on specific FEMNIST and CIFAR tasks.

Communication Efficiency. Table 2 highlights HaFedHo's accelerated convergence. By ensuring that local objective functions remain consistent with the global objective, HaFedHo aggressively dampens the client-drift caused by non-IID data, requiring significantly fewer communication rounds to reach target accuracy thresholds compared to prior arts.

Robustness to Data Heterogeneity. As illustrated in Table 1 and Fig. 1, heightened statistical heterogeneity uniformly degrades model performance across all frameworks. However, HaFedHo dominates across all non-IID regimes. This resilience is achieved by cross-pollinating objective function data among clients, ensuring local update directions are tightly aligned with the ideal global gradient.

Resilience to Low Participation. Production FL systems typically sample minute fractions from vast device pools [Shejwalkar *et al.*, 2022], rendering the conventional 10% participation assumption [Wang *et al.*, 2020b; Acar *et al.*, 2021] unrealistic. Table 3 evaluates extreme low-participation regimes (0.2%, 0.5%, 1%). In the highly skewed CIFAR-10 setting, where competitors suffer catastrophic performance drops, HaFedHo maintains robust accuracy trajectories well above the FedAvg baseline.

Scalability to Massive Cohorts. We simulate large-scale parallel updates on FEMNIST using 500 and 1020 active clients (Figs. 2a and 2b). Most existing methods fail to scale, matching basic FedAvg. Crucially, SCAFFOLD diverges entirely, corroborating prior findings [Li *et al.*, 2021; Karimireddy *et al.*, 2021; Karimireddy *et al.*, 2020; Khanduri *et al.*, 2021] that strictly bounded control variates fail under massive decentralization. HaFedHo bypasses this bottleneck, effectively absorbing global feedback to refine local objectives and achieving superior convergence rates.

5.3 Ablation Study

The core of HaFedHo is its *objective rectification* strategy that rectifies the local objective to be $f_i(\boldsymbol{w}) + \mu\Big(\hat{f} + \boldsymbol{h}^\top(\boldsymbol{w} - \boldsymbol{w}^t)\Big)$. A simpler alternative might be to perform *gradient rectification*, where the correction term $\mu\boldsymbol{h}$ is added directly to the local gradient $\nabla f_i(\boldsymbol{w})$ instead of being incorporated into the objective function via its integral. We compare these two strategies in Figure 3.

The figure displays the accuracy curves of the two methods under different μ over CIFAR-10 and CIFAR-100. By tuning μ to the best value, gradient rectification (the best $\mu = 0.001$ for CIFAR-10 and CIFAR-100) only marginally outperforms the FedAvg baseline, whereas our objective rectification (the best $\mu = 0.01$ for CIFAR-10 and CIFAR-100) substantially enhances FL. This demonstrates that simply correcting the gradient direction is insufficient. By modifying the objective function itself, HaFedHo ensures that the entire local optimization trajectory at the client is better aligned with the centralized training goal, leading to a much more effective mitigation of client drift.

6 Limitations and Future Work

Higher-Order Taylor Approximations. Currently, HaFedHo aligns local objectives with the centralized ideal using a first-order Taylor series expansion. Although this approach yields substantial performance gains in federated settings, a persistent accuracy gap remains when compared to purely centralized training. This discrepancy primarily stems from the approximation errors inherent in first-order truncations. Incorporating higher-order Taylor terms could

Method	10% sampling rate, FEMNIST (3400 clients), CIFAR10 (100 clients), CIFAR100 (100 clients)										
	FEMNIST, CNN	CIFAR10, ResNet18				CIFAR100, ResNet18					
	Non-IID	q=0.3	q=0.4	q=0.6	IID	q=0.02	q=0.1	q=0.3	q=0.4	q=0.6	IID
Centralized Learning	**84.39**$_{(\pm 0.14)}$	**85.01**$_{(\pm 0.21)}$	**85.24**$_{(\pm 0.15)}$	**84.98**$_{(\pm 0.23)}$	**85.29**$_{(\pm 0.12)}$	**48.81**$_{(\pm 0.24)}$	**49.09**$_{(\pm 0.19)}$	**48.79**$_{(\pm 0.22)}$	**49.11**$_{(\pm 0.16)}$	**48.85**$_{(\pm 0.25)}$	**49.07**$_{(\pm 0.20)}$
FedAvg	78.53$_{(\pm 0.32)}$	72.38$_{(\pm 0.25)}$	73.27$_{(\pm 0.41)}$	75.84$_{(\pm 0.18)}$	80.39$_{(\pm 0.39)}$	27.38$_{(\pm 0.26)}$	33.97$_{(\pm 0.31)}$	39.29$_{(\pm 0.22)}$	39.33$_{(\pm 0.48)}$	41.25$_{(\pm 0.34)}$	41.52$_{(\pm 0.29)}$
HaFedHo	81.25$_{(\pm 0.24)}$	76.20$_{(\pm 0.18)}$	77.01$_{(\pm 0.29)}$	79.88$_{(\pm 0.33)}$	82.21$_{(\pm 0.15)}$	28.42$_{(\pm 0.21)}$	37.52$_{(\pm 0.42)}$	42.36$_{(\pm 0.16)}$	43.89$_{(\pm 0.27)}$	45.10$_{(\pm 0.31)}$	46.59$_{(\pm 0.22)}$
FedProx	78.41$_{(\pm 0.41)}$	73.69$_{(\pm 0.19)}$	74.81$_{(\pm 0.32)}$	76.79$_{(\pm 0.34)}$	79.98$_{(\pm 0.22)}$	18.59$_{(\pm 0.37)}$	34.11$_{(\pm 0.28)}$	39.71$_{(\pm 0.31)}$	37.49$_{(\pm 0.24)}$	38.22$_{(\pm 0.45)}$	41.91$_{(\pm 0.16)}$
SCAFFOLD	79.29$_{(\pm 0.27)}$	70.52$_{(\pm 0.42)}$	72.36$_{(\pm 0.19)}$	75.48$_{(\pm 0.27)}$	78.59$_{(\pm 0.43)}$	1.15$_{(\pm 0.21)}$	13.54$_{(\pm 0.39)}$	34.56$_{(\pm 0.23)}$	34.85$_{(\pm 0.41)}$	37.58$_{(\pm 0.16)}$	38.57$_{(\pm 0.36)}$
MOON	79.11$_{(\pm 0.31)}$	71.84$_{(\pm 0.16)}$	73.51$_{(\pm 0.45)}$	75.63$_{(\pm 0.21)}$	79.89$_{(\pm 0.35)}$	26.21$_{(\pm 0.21)}$	34.28$_{(\pm 0.43)}$	38.69$_{(\pm 0.18)}$	40.27$_{(\pm 0.33)}$	41.19$_{(\pm 0.17)}$	43.14$_{(\pm 0.47)}$
MimeLite	78.82$_{(\pm 0.23)}$	74.39$_{(\pm 0.38)}$	75.39$_{(\pm 0.27)}$	77.96$_{(\pm 0.41)}$	81.08$_{(\pm 0.42)}$	27.92$_{(\pm 0.18)}$	34.19$_{(\pm 0.33)}$	40.13$_{(\pm 0.34)}$	41.59$_{(\pm 0.22)}$	41.54$_{(\pm 0.47)}$	41.63$_{(\pm 0.29)}$
FedDyn	80.29$_{(\pm 0.33)}$	58.85$_{(\pm 0.28)}$	68.97$_{(\pm 0.16)}$	77.49$_{(\pm 0.44)}$	82.01$_{(\pm 0.31)}$	3.91$_{(\pm 0.42)}$	3.89$_{(\pm 0.19)}$	33.19$_{(\pm 0.47)}$	37.53$_{(\pm 0.25)}$	43.21$_{(\pm 0.36)}$	46.41$_{(\pm 0.41)}$
FedLWS	80.95$_{(\pm 0.25)}$	74.99$_{(\pm 0.41)}$	75.89$_{(\pm 0.22)}$	78.91$_{(\pm 0.33)}$	82.03$_{(\pm 0.34)}$	27.81$_{(\pm 0.25)}$	34.95$_{(\pm 0.46)}$	41.22$_{(\pm 0.28)}$	42.91$_{(\pm 0.31)}$	44.32$_{(\pm 0.19)}$	44.71$_{(\pm 0.44)}$
pFedAFM	79.79$_{(\pm 0.26)}$	74.02$_{(\pm 0.31)}$	75.19$_{(\pm 0.18)}$	78.77$_{(\pm 0.42)}$	81.83$_{(\pm 0.23)}$	28.14$_{(\pm 0.31)}$	34.71$_{(\pm 0.15)}$	39.89$_{(\pm 0.49)}$	41.54$_{(\pm 0.27)}$	43.08$_{(\pm 0.32)}$	44.59$_{(\pm 0.19)}$

Table 1: Validation accuracy (%) with a 10% client sampling rate across non-IID settings controlled by the Dirichlet hyperparameter q. Results show the mean and standard deviation across five randomized initializations. Smaller q values correspond to higher degrees of statistical heterogeneity. Optimal performance is indicated in **boldface**, while runner-up methods are highlighted in navy.

Framework	**FEMNIST (CNN)**		**CIFAR-10 (ResNet18)**		**CIFAR-100 (ResNet18)**	
	Rounds	Convergence Rate	Rounds	Convergence Rate	Rounds	Convergence Rate
FedAvg	312	1.00×	894	1.00×	741	1.00×
	942	1.00×	1922	1.00×	1712	1.00×
FedProx	355	0.88×	841	1.06×	962	0.77×
	892	1.06×	1980	0.97×	2311	0.74×
SCAFFOLD	321	0.97×	1045	0.86×	1682	0.44×
	822	1.15×	1935	0.99×	2500+	<0.68×
MimeLite	288	1.08×	422	2.12×	551	1.34×
	850	1.11×	1361	1.41×	1410	1.21×
FedDyn	291	1.07×	1852	0.48×	2500+	<0.30×
	628	1.50×	2500+	<0.77×	2500+	<0.68×
MOON	310	1.01×	811	1.10×	822	0.90×
	861	1.09×	1910	1.01×	1792	0.96×
FedLWS	165	1.89×	410	2.18×	533	1.39×
	382	2.47×	995	1.93×	1194	1.43×
pFedAFM	241	1.29×	442	2.02×	540	1.37×
	575	1.64×	1110	1.73×	1342	1.28×
HaFedHo (Ours)	**122**	**2.56**×	**371**	**2.41**×	**595**	**1.25**×
	284	**3.32**×	**1052**	**1.83**×	**1310**	**1.31**×

Table 2: Evaluation of network training efficiency. The metrics capture the exact quantity of global epochs required to reach targeted accuracy metrics. For the FEMNIST setup, evaluations are benchmarked at 69% and 79% accuracy thresholds; for CIFAR-10, the limits are set at 65% and 75%; while for CIFAR-100, they follow 30% and 40%. Entries designated with 2500+ represent runs failing to achieve convergence parameters. Speedup performance indicators are scaled against the standardized FedAvg pipeline.

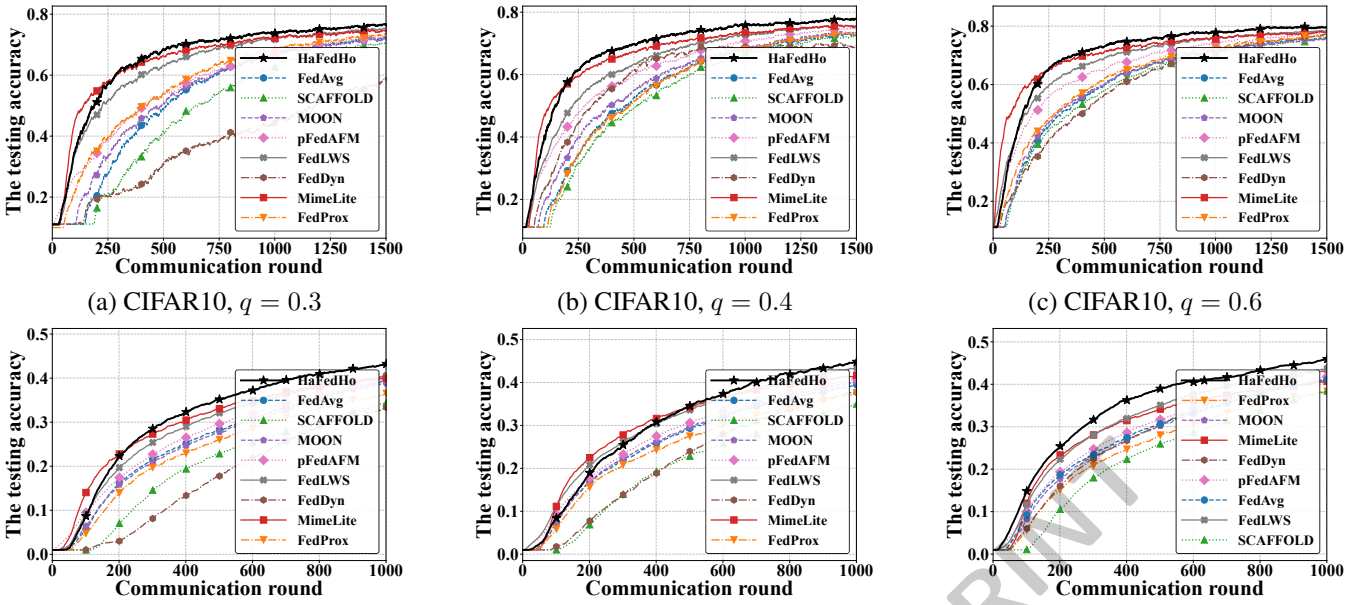
Framework	**FEMNIST, CNN, 3400 clients**			**CIFAR10, ResNet18, 1000 clients**		
	Participation Levels			Participation Levels		
	0.20%	0.50%	1%	0.20%	0.50%	1%
Centralized	**84.45**$_{(\pm 0.11)}$	**84.45**$_{(\pm 0.11)}$	**84.45**$_{(\pm 0.11)}$	**85.19**$_{(\pm 0.15)}$	**85.19**$_{(\pm 0.15)}$	**85.19**$_{(\pm 0.15)}$
FedAvg	78.81$_{(\pm 0.18)}$	79.22$_{(\pm 0.21)}$	79.51$_{(\pm 0.32)}$	70.62$_{(\pm 0.38)}$	74.15$_{(\pm 0.28)}$	75.77$_{(\pm 0.22)}$
FedProx	78.48$_{(\pm 0.11)}$	79.41$_{(\pm 0.33)}$	79.75$_{(\pm 0.45)}$	50.15$_{(\pm 0.22)}$	63.42$_{(\pm 0.35)}$	67.81$_{(\pm 0.49)}$
SCAFFOLD	78.65$_{(\pm 0.25)}$	79.68$_{(\pm 0.14)}$	79.88$_{(\pm 0.38)}$	70.82$_{(\pm 0.44)}$	76.05$_{(\pm 0.26)}$	80.25$_{(\pm 0.41)}$
MOON	78.75$_{(\pm 0.39)}$	78.85$_{(\pm 0.45)}$	79.12$_{(\pm 0.19)}$	70.35$_{(\pm 0.52)}$	72.15$_{(\pm 0.21)}$	73.85$_{(\pm 0.55)}$
MimeLite	78.33$_{(\pm 0.48)}$	78.95$_{(\pm 0.31)}$	79.85$_{(\pm 0.64)}$	53.85$_{(\pm 0.51)}$	70.75$_{(\pm 0.48)}$	75.33$_{(\pm 0.18)}$
FedDyn	79.65$_{(\pm 0.42)}$	80.12$_{(\pm 0.18)}$	80.75$_{(\pm 0.49)}$	–	–	–
FedLWS	78.75$_{(\pm 0.25)}$	79.48$_{(\pm 0.36)}$	80.18$_{(\pm 0.39)}$	70.45$_{(\pm 0.41)}$	75.88$_{(\pm 0.25)}$	80.65$_{(\pm 0.34)}$
pFedAFM	78.58$_{(\pm 0.29)}$	79.35$_{(\pm 0.24)}$	79.91$_{(\pm 0.44)}$	55.35$_{(\pm 0.48)}$	71.22$_{(\pm 0.31)}$	76.45$_{(\pm 0.42)}$
HaFedHo (Ours)	80.12$_{(\pm 0.27)}$	80.65$_{(\pm 0.31)}$	81.28$_{(\pm 0.22)}$	71.35$_{(\pm 0.33)}$	76.88$_{(\pm 0.29)}$	81.35$_{(\pm 0.46)}$

Table 3: Comparative analysis of top-1 test accuracy (%) across varying client sampling ratios. The optimal theoretical baseline (Centralized) is denoted in **bold**, while the leading federated framework is highlighted in navy. Hyphens indicate scenarios where computational limitations were exceeded.

theoretically provide a more precise objective rectification. However, doing so introduces a critical trade-off between computational feasibility and estimation accuracy. Exploring lightweight, higher-order approximation techniques remains a promising avenue for future research.

7 Conclusion

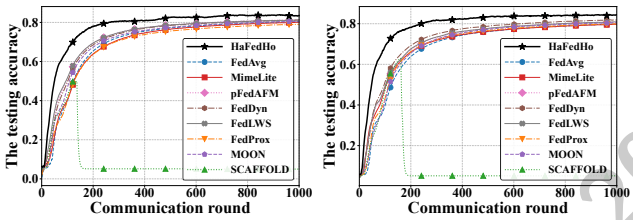
We presented HaFedHo, a novel FL framework that bridges the gap between local client updates and the global data-centralized optimum. We prove that the inherent estimation error naturally decays over the course of training. Comprehensive empirical evaluations validate our theoretical claims, demonstrating that HaFedHo achieves superior terminal accuracy and accelerated communication efficiency compared to existing state-of-the-art baselines. Furthermore, it exhibits exceptional resilience under highly restrictive real-world con-



(a) CIFAR10, $q = 0.3$ (b) CIFAR10, $q = 0.4$ (c) CIFAR10, $q = 0.6$

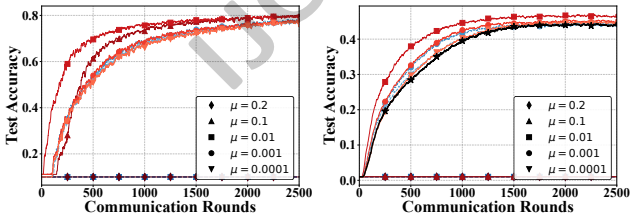
(d) CIFAR100, $q = 0.3$ (e) CIFAR100, $q = 0.4$ (f) CIFAR100, $q = 0.6$

Figure 1: The impact of non-IID degrees on the test accuracy of global model over CIFAR10 and CIFAR100.



(a) 500 clients. (b) 1020 clients.

Figure 2: Evaluation of scalability under massive client participation on the FEMNIST dataset. Benefiting from its objective rectification, HaFedHo maintains robust convergence and scales seamlessly as the cohort size expands. In contrast, the control variates of SCAFFOLD deteriorate, leading to divergence as previously observed in the literature [Karimireddy *et al.*, 2021].



(a) Ablation on CIFAR-10. (b) Ablation on CIFAR-100.

Figure 3: Ablation study comparing Objective Rectification (our method with red line) with a simpler gradient rectification approach (blue line). Modifying the objective function is crucial for performance.

ditions, including severe data heterogeneity and extremely low client participation regimes.

Acknowledgments

This work was supported by the Key R&D Program of Zhejiang Province under Grant No. 2025C01084 and the CAAI–MindSpore Open Fund, developed on the OpenI Community, under Contract No. CAAIXSJLJJ 2025 MindSpore 04.

References

[Acar *et al.*, 2021] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

[Bakhtiari *et al.*, 2020] Mohammad Bakhtiari, Reza Nasirigerdeh, Reihaneh Torkzadehmahani, Amirhossein Bayat, David B Blumenthal, Markus List, and Jan Baumbach. Federated multi-mini-batch: An efficient training approach to federated learning in non-iid environments [j]. *arXiv preprint arXiv:2011.07006*, 2020.

[Caldas *et al.*, 2018] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[Chen and Chao, 2020] Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2020.

[Chen *et al.*, 2024] Jinqian Chen, Haoyu Tang, Junhao Cheng, Ming Yan, Ji Zhang, Mingzhu Xu, Yupeng Hu,

and Liqiang Nie. Breaking barriers of system heterogeneity: straggler-tolerant multimodal federated learning via knowledge distillation. In *Ihe International Joint Conference on Artificial Intelligence*, 2024.

[Chu *et al.*, 2025] Yun-Wei Chu, Dong-Jun Han, Seyyedali Hosseinalipour, and Christopher Brinton. Unlocking the potential of model calibration in federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Gao *et al.*, 2022] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10112–10121, 2022.

[Geiping *et al.*, 2020] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[He *et al.*,] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*.

[Hsu *et al.*, 2019] 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.

[Hu *et al.*, 2024] Ke Hu, Liyao Xiang, Peng Tang, and Weidong Qiu. Feature norm regularized federated learning: utilizing data disparities for model performance gains. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4136–4146, 2024.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[Karimireddy *et al.*, 2021] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[Khanduri *et al.*, 2021] Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061, 2021.

[Li *et al.*, 2020a] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[Li *et al.*, 2020b] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2020.

[Li *et al.*, 2021] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021.

[Lu *et al.*, 2023] Jianrong Lu, Lulu Xue, Wei Wan, Minghui Li, Leo Yu Zhang, and Shengqing Hu. Preserving privacy of input features across all stages of collaborative learning. In *2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 191–198. IEEE, 2023.

[Lu *et al.*, 2024] Jianrong Lu, Shengshan Hu, Wei Wan, Minghui Li, Leo Yu Zhang, Lulu Xue, and Hai Jin. Depriving the survival space of adversaries against poisoned gradients in federated learning. *IEEE Transactions on Information Forensics and Security*, 19:5405–5418, 2024.

[Lu *et al.*, 2025] Jianrong Lu, Zhiyu Zhu, and Junhui Hou. Parasolver: A hierarchical parallel integral solver for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lu *et al.*, 2026] Jianrong Lu, Junwei Liu, Xingyun Zheng, Minghui Yang, Jian Wang, Ping Wang, and Yechao Zhang. Mhb: Medical hallucination benchmark for large language models in complex clinical tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38971–38978, 2026.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, volume 54, pages 1273–1282, 2017.

[Ning *et al.*, 2024] Zhiyuan Ning, Chunlin Tian, Meng Xiao, Wei Fan, Pengyang Wang, Li Li, Pengfei Wang, and Yuanchun Zhou. Fedgcs: A generative framework for efficient client selection in federated learning via gradient-based optimization. *The Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

[Pathak and Wainwright, 2020] Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in Neural Information Processing Systems*, 33:7057–7066, 2020.

[Shejwalkar *et al.*, 2022] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of

poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1354–1371. IEEE, 2022.

[Shen *et al.*, 2025] Lei Shen, Zhenheng Tang, Lijun Wu, Yonggang Zhang, Xiaowen Chu, Tao Qin, and Bo Han. Hot-pluggable federated learning: Bridging general and personalized FL via dynamic selection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Shi *et al.*, 2025a] Changlong Shi, Jinmeng Li, He Zhao, Dan dan Guo, and Yi Chang. FedLWS: Federated learning with adaptive layer-wise weight shrinking. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Shi *et al.*, 2025b] Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30651–30660, 2025.

[Sun *et al.*, 2023] Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. *arXiv preprint arXiv:2305.11584*, 2023.

[Sun *et al.*, 2025] Zhenyu Sun, ziyang zhang, Zheng Xu, Gauri Joshi, Pranay Sharma, and Ermin Wei. Debiasing federated learning with correlated client participation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Wang *et al.*, 2020a] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020.

[Wang *et al.*, 2020b] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[Wang *et al.*, 2023] Shuai Wang, Yanqing Xu, Zhiguo Wang, Tsung-Hui Chang, Tony QS Quek, and Defeng Sun. Beyond admm: A unified client-variance-reduced adaptive federated learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10175–10183, 2023.

[Wang *et al.*, 2024] Jiaqi Wang, Xingyi Yang, Suhan Cui, Liwei Che, Lingjuan Lyu, Dongkuan DK Xu, and Fenglong Ma. Towards personalized federated learning via heterogeneous model reassembly. *Advances in Neural Information Processing Systems*, 36, 2024.

[Xing *et al.*, 2025] X. Xing, Q. Zhan, X. Xie, et al. Flexible sharpness-aware personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21707–21715, 2025.

[Xu *et al.*, 2021] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[Xu *et al.*, 2024] Yi Xu, Ying Li, Haoyu Luo, Xiaoliang Fan, and Xiao Liu. Fblg: A local graph based approach for handling dual skewed non-iid data in federated learning. In *The Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

[Yi *et al.*, 2025] Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. pFedAFM: Adaptive Feature Mixture for Data-Level Personalization in Heterogeneous Federated Learning on Mobile Edge Devices . In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 1981–1994, Los Alamitos, CA, USA, May 2025. IEEE Computer Society.

[Yoon *et al.*, 2021] Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv preprint arXiv:2107.00233*, 2021.

[Zeng *et al.*, 2024] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Zhiqi Shen, Hanchao Yu, and Yingwei Zhang. Fedes: federated early-stopping for hindering memorizing heterogeneous label noise. In *The Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5416–5424, 2024.

[Zheng *et al.*, 2025] Hao Zheng, Zhigang Hu, Liu Yang, Meiguang Zheng, Aikun Xu, and Boyu Wang. Fedcalm: Conflict-aware layer-wise mitigation for selective aggregation in deeper personalized federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15444–15453, 2025.