Utility-Driven Clustered Federated Learning via Class-wise Interaction Attribution

Zhongxu Wang[1], Yichen Li[1], Haozhao Wang[1], Hanlin Cai[2], Qi Chen[4]
Ruixuan Li[1]*, Ozgur B. Akan[2,3]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China [2]Centre for neXt Communications (CXC), Department of Engineering, University of Cambridge, Cambridge, U.K.[3] Centre for neXt Communications (CXC), Department of Electrical and Electronics Engineering, Koç University, Istanbul, Türkiye.
[4]Wuhan Dameng Database Co., Ltd, Wuhan, China
{wzx11,ycli0204,hz_wang}@hust.edu.cn

Abstract

Clustered Federated Learning (CFL) mitigates statistical heterogeneity by grouping clients with related learning tasks. However, existing CFL frameworks predominantly rely on proxy metrics, such as gradient or data similarities, which collapse fine-grained class-level similarity into macroscopic client-level scores. This coarse-grained aggregation inherently masks critical class-wise conflicts, inadvertently exposing the federation to severe negative transfer. In this paper, we propose a fundamental paradigm shift from heuristic similarity to *explicit value attribution*. We introduce **FedCap**, a novel utility-driven CFL framework that explicitly quantifies class-wise marginal utilities to capture the true generalization impact of local updates. Guided by this attribution mechanism, FedCap employs a conflict-aware clustering algorithm to systematically eliminate intra-cluster negative transfer, while orchestrating safe, supply-demand-guided representation sharing across clusters. Crucially, we provide rigorous theoretical guarantees demonstrating the robustness of our underlying attribution mechanism against four major types of distribution shifts. Extensive experiments across diverse heterogeneity settings show that FedCap consistently achieves state-of-the-art accuracy.

1 Introduction

Federated Learning (FL) has established itself as the prevailing paradigm for collaboratively training neural networks across decentralized data silos while preserving data privacy [39, 31, 34, 32, 33]. However, traditional FL frameworks remain fundamentally bottlenecked by statistical heterogeneity [36, 28, 26]. In practice, highly Non-IID local data typically manifests in complex distribution shifts, notably label skew [59], feature skew [24], quantity shift [52], and concept shift [18](see Appx. A.1 for definitions).

To overcome this dilemma, Clustered FL (CFL) has emerged as a promising solution that groups clients into distinct clusters based on task affinities, thereby maintaining a specialized global model for each cluster. Within this paradigm, data-centric approaches [50] partition clients by computing principal angles between local data subspaces, whereas gradient-based methods [46, 61] evaluate affinities through the similarity of local model updates. Alternative model-centric strategies iteratively assign clients to K concurrently maintained global models [12, 11, 37]. More recently, to simultaneously tackle multiple distribution shifts, FedDAG [43] integrates both data and gradient

*Ruixuan Li is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Clients \ Classes	Airplane	Auto	Truck	Bird
T1	20	300	300	0
T2	20	300	300	400
T3	400	300	0	300

Figure 1: Client data distributions. Client *T1* lacks enough "Airplane" and shares majority classes with client *T2*, whereas client *T3* complements *T1*.

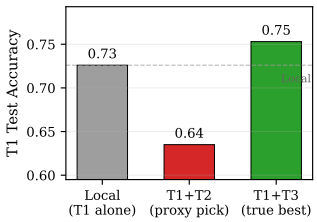

Figure 2: Similarity proxy metrics pick client *T2*, but client *T3* is the true helper.

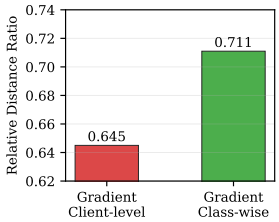

Figure 3: Coarse-grained client-level similarity masks class-level negative transfer.

information into a hybrid similarity metric. However, effectively grouping heterogeneous clients using existing methods remains sub-optimal due to the following limitations:

Limit 1. Similarity Does Not Necessarily Predict Transfer Utility. Existing CFL approaches use gradient [46, 61] or data-subspace [50] similarities to group clients. These descriptors do not directly evaluate whether a local update reduces loss on a shared reference task. Data similarity describes input structure, while gradient similarity depends on the model and optimization state. Neither alone guarantees beneficial collaboration. For example, Fig. 2 shows that jointly training the statistically more similar clients *T1* and *T2* can reduce accuracy, whereas collaboration between *T1* and the more dissimilar *T3* improves it. This motivates evaluating class-wise transfer utility relative to a common reference set.

Limit 2. Client-Level Aggregation Can Mask Class-Wise Conflicts. Aggregating interactions into a client-level scalar can allow majority-class overlap to mask conflicts affecting minority classes. To illustrate this effect, we evaluate the relative distance ratio $d(T1,T2)/d(T1,T3)$. Ratios below one rank *T2* closer than *T3*, despite the latter's greater benefit to *T1* in this example. As shown in Fig. 3, the traditional gradient metric yields a low ratio of **0.645**, showing a strong misleading bias towards *T2*. In contrast, evaluating cosine similarity between gradients derived exclusively from samples of shared classes increases this ratio to **0.711**. This example illustrates how class-level information changes the similarity estimate. Since both ratios remain below one, restricting gradients to shared classes alone does not correct the ranking.

These limitations raise the following crucial **question**: *Can we formulate a utility-driven attribution metric that explicitly captures fine-grained class-to-class interactions, thereby guaranteeing mathematically robust clustering and avoid cross-class negative transfer?*

Our contributions. We introduce class-wise utility attribution as a clustering criterion. Clients with aligned predictive biases can exhibit comparable interaction profiles on a shared validation set under the conditions analyzed below. In **FedCap**, each client estimates these profiles during warm-up using a small, class-balanced public validation set. The server uses them to penalize within-cluster conflicts and select complementary clusters for encoder sharing. Our contributions are:

- We propose the Class-interaction Attribution Matrix (**CAM**) to estimate class-to-class utility. We analyze its stability under label and quantity shifts and bounded feature skew, and its raw-row separability under concept shift, subject to explicit assumptions.
- We design **FedCap**, which uses CAM to penalize cross-class conflicts during clustering and guide selective encoder sharing across clusters.
- Experiments show higher mean accuracy than the compared baselines across the evaluated settings. A server-side timing study shows lower clustering cost than the evaluated gradient, data-subspace, and fusion descriptors.

2 Preliminaries and Problem Statement

2.1 In-Run Data Shapley

In this subsection, we briefly review the formal definition of Shapley and introduce the concept of In-Run Data Shapley, which enables efficient data attribution during the model training process.

Definition 1 (*Data Shapley in Machine Learning [10]*). Consider a training dataset $\mathcal{D}_{train} = \{z_1, z_2, \ldots, z_N\}$ and a clean validation set $\mathcal{D}_{val}$. Let $V(S)$ denote a utility function (e.g., model

accuracy or negative validation loss) evaluated on $\mathcal{D}_{val}$ using a model trained exclusively on a data subset $S \subseteq \mathcal{D}_{train}$. Data Shapley of a specific data point z_i is defined as:

$$\phi_i(V) = \sum_{S \subseteq \mathcal{D}_{train} \setminus \{z_i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left[V(S \cup \{z_i\}) - V(S) \right], \tag{1}$$

where $|\cdot|$ denotes set cardinality. Evaluating this exact formulation requires retraining the model for all 2^N possible subsets, incurring a prohibitive exponential computational cost. Despite this intractability, the Shapley value remains the gold standard for data valuation because it is the *unique* metric mathematically guaranteed to satisfy four fundamental fairness axioms. While comprehensive definitions of all axioms are deferred to Appx. A.3, we highlight the *Linearity* and *Efficiency* axioms below, as they provide the theoretical backbone for our proposed class-wise aggregation mechanism.

Axiom 1 (*Linearity*). For any two utility functions V_1, V_2 and real scalars $\alpha, \beta \in \mathbb{R}$, the attribution is strictly linear: $\phi_i(\alpha V_1 + \beta V_2) = \alpha\phi_i(V_1) + \beta\phi_i(V_2)$.

Axiom 2 (*Efficiency*). The total utility generated by the entire dataset is exactly and exhaustively distributed among all individual data points: $\sum_{i=1}^{N} \phi_i(V) = V(\mathcal{D}_{train}) - V(\emptyset)$.

In-Run Data Shapley via Ghost Dot-Product. To avoid the prohibitive computational cost of retraining and quantify data contributions to a *specific* trained model, In-Run Data Shapley (IRDS) [51] introduces a highly efficient approximation that estimates data valuation within a single training trajectory. Let θ_t represent the model parameters at training step t, and η_t be the learning rate. The parameter update induced by a single data point z_i can be approximated as $\Delta\theta_t^{(i)} = -\eta_t \nabla_\theta \mathcal{L}(z_i, \theta_t)$, where $\mathcal{L}$ is the training loss. By leveraging the first-order Taylor expansion, the instantaneous marginal utility contributed by z_i at step t can be quantified tightly by the inner product:

$$u_t(i) := -\nabla_\theta \mathcal{L}(z^{\text{val}}, \theta_t)^\top \Delta\theta_t^{(i)} = \eta_t \langle \nabla_\theta \mathcal{L}(z^{\text{val}}, \theta_t), \nabla_\theta \mathcal{L}(z_i, \theta_t) \rangle. \tag{2}$$

Here, $u_t(i)$ estimates the validation-loss reduction due to the update from z_i: $\mathcal{L}(z^{\text{val}}, \theta_t) - \mathcal{L}(z^{\text{val}}, \theta_t + \Delta\theta_t^{(i)}) \approx u_t(i)$. Positive values indicate estimated benefit. By Linearity, the Shapley values of the additive first-order step utilities accumulate as $\hat{\phi}_i = \sum_{t=1}^{T} u_t(i)$. This is a first-order in-run attribution estimate, rather than an exact retraining-based Data Shapley value. To further alleviate the computational overhead, IRDS [51] leverages the *ghost dot-product* technique. Specifically, for a neural network layer with weight $\mathbf{W}$, input activation $\mathbf{a}^{(i)}$, and individual's output gradient $\mathbf{e} = \frac{\partial \ell^{(i)}}{\partial \mathbf{s}^{(i)}} = \frac{\partial \ell}{\partial \mathbf{s}^{(i)}}$, the inner product of weight gradients between a validation sample and a training sample is algebraically decomposed via the chain rule:

$$\left\langle \frac{\partial \ell_{val}}{\partial \mathbf{W}}, \frac{\partial \ell_{train}}{\partial \mathbf{W}} \right\rangle = \left(\mathbf{a}_{val}^\top \mathbf{a}_{train} \right) \left(\mathbf{e}_{val}^\top \mathbf{e}_{train} \right). \tag{3}$$

This decomposition computes the gradient dot-product without materializing per-sample parameter-gradient vectors. It still requires validation-sample processing and activation/error inner products. We aggregate these sample-level attribution estimates into *class-to-class interactions*.

2.2 Problem Statement

Consider a FL system with N clients, where each client i possesses a local private dataset $\mathcal{D}_i$. Under severe statistical heterogeneity, forcing all clients to converge to a single global model leads to destructive interference. To address this, CFL partitions the N clients into K disjoint clusters $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$, where $\bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \ldots, N\}$. The CFL objective jointly optimizes the cluster assignments $\mathcal{C}$ and a set of specialized cluster models $\{w_1, \ldots, w_K\}$:

$$\min_{\mathcal{C}, \{w_k\}} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \frac{|\mathcal{D}_i|}{|\mathcal{D}_{\mathcal{C}_k}|} \mathcal{L}_i(w_k), \text{ where } \mathcal{L}_i(w_k) = \frac{1}{|\mathcal{D}_i|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_i} \mathcal{L}_{CE}\big(\phi_k(\theta_k(\mathbf{x})), y\big). \tag{4}$$

where $\mathcal{D}_{\mathcal{C}_k} = \bigcup_{i \in \mathcal{C}_k} \mathcal{D}_i$ represents the total data volume within cluster k. To facilitate flexible knowledge sharing, we decouple the model parameters w_k into a feature encoder θ_k and a personalized classifier head ϕ_k, i.e., $w_k = (\theta_k, \phi_k)$. Solving Eq. 4 fundamentally depends on the accurate evaluation of client affinities to form optimal clusters $\mathcal{C}$ and the effective collaboration among them.

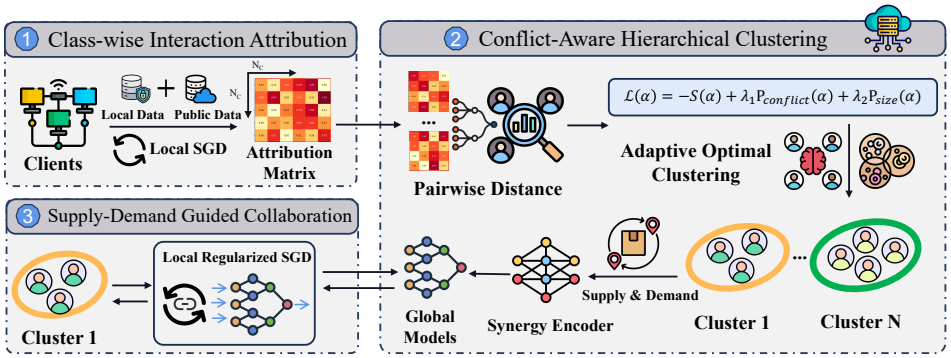

Figure 4: FedCap workflow: warm-up CAM estimation, conflict-aware clustering, and cross-cluster encoder alignment. Heads are shared within clusters only.

3 Methodology

In this section, we introduce the *Class-interaction Attribution Matrix (CAM)*, derived from first-order IRDS, and analyze its behavior under four types of distribution shift under explicit assumptions. Subsequently, we propose the FedCap framework. By leveraging CAM, FedCap executes a novel hierarchical clustering algorithm guided by a customized loss function that explicitly mitigates intra-cluster class conflicts, followed by an adaptive cross-cluster representation sharing mechanism. The comprehensive workflow is detailed in Alg. 1, and Fig. 4 illustrates the overall architecture.

3.1 Class-wise Interaction Attribution via Local Training

While the original IRDS [51] quantifies the value of individual data points, such granular metrics are computationally redundant for identifying statistical heterogeneity in FL. Fundamentally, complex distribution shifts rarely occur as isolated point-wise anomalies, but rather manifest as *class-level* phenomena across multiple samples. Therefore, we propose to elevate the sample-level attribution into a class-wise paradigm by constructing local *Class-interaction Attribution Matrix (CAM)*.

Prior to training, the server broadcasts a small, labeled, class-balanced validation set $\mathcal{D}_{val}^{pub}$ as a shared reference. Class balance gives classes equal reference coverage; it does not guarantee representativeness across client domains. At each warm-up step t, client i concatenates a local training mini-batch $\mathcal{B}_{tr}$ and a validation mini-batch $\mathcal{B}_{val}$ for attribution computation. The parameter update uses the local training gradients; validation samples provide reference gradients. Utilizing the *ghost dot-product* technique, the client efficiently computes the point-to-point instantaneous utility $\Phi_{\mathbf{x}\to\mathbf{x}'}^{(t)} = \eta_t \langle \nabla_\theta \mathcal{L}_{val}(\mathbf{x}'), \nabla_\theta \mathcal{L}_{tr}(\mathbf{x}) \rangle$ between any training sample $\mathbf{x}$ and validation sample $\mathbf{x}'$. We aggregate these first-order attribution estimates by validation class and training class:

- **Validation-side:** For a utility defined as the sum of validation-sample utilities, Linearity allows aggregation over reference samples of class c': $\Phi_{\mathbf{x}\to c'}^{(t)} = \sum_{\mathbf{x}'\in\mathcal{B}_{val}^{c'}} \Phi_{\mathbf{x}\to\mathbf{x}'}^{(t)}$.
- **Training-side:** For the additive first-order surrogate, class-wise attribution is the sum over local training samples labeled c.

Consequently, by accumulating these cross-class interactions over the T-step warm-up trajectory, the element at row c and column c' of the CAM for client i is rigorously formulated as:

$$\mathbf{M}_{c,c'}^{(i)} = \sum_{t=1}^{T} \sum_{\mathbf{x}\in\mathcal{B}_{tr}^{c}} \sum_{\mathbf{x}'\in\mathcal{B}_{val}^{c'}} \eta_t \big\langle \nabla_\theta \mathcal{L}_{val}(\mathbf{x}', \theta_t), \nabla_\theta \mathcal{L}_{tr}(\mathbf{x}, \theta_t) \big\rangle. \tag{5}$$

The sign estimates whether class-c updates increase (< 0) or reduce (> 0) the loss on reference class c' during warm-up. The resulting $C \times C$ matrix $\mathbf{M}^{(i)}$ serves as a multi-dimensional geometric signature, encapsulating both the local data distribution and the model's capability. To theoretically ground this heuristic, we bridge the CAM with the Neural Tangent Kernel formulation of gradient

dot-products. Let $\mathbf{v}_c^{(i)}$ denote the L_2-normalized c-th row of client i's CAM. We establish three rigorous theoretical guarantees regarding its topological behavior under diverse distribution shifts:

Theorem 1 (Isomorphism under Label Skew and Quantity Shift). *For clients i and j without concept shift, assume a shared predictive-bias regime, bounded residual and kernel mismatch, and the concentration conditions in Appx. A.4. For shared classes with nonzero CAM rows, with probability at least $1-\delta$, $\langle \mathbf{v}_c^{(i)}, \mathbf{v}_c^{(j)} \rangle \geq 1 - \epsilon_{LQ}(c, T, \delta)$. The error depends on residual mismatch, kernel mismatch, and empirical fluctuations.*

Theorem 2 (Robustness under Feature Skew). *Given a bounded feature perturbation ε_f on the class-conditional NTK, the CAM topology remains stable relative to a normal reference client 0, yielding $\langle \mathbf{v}_c^{(i)}, \mathbf{v}_c^{(0)} \rangle \geq 1 - \epsilon_F(c, T, \delta)$. Furthermore, the diagonal dominance of the CAM is strictly preserved provided that the semantic NTK margin exceeds the perturbation level ($\kappa_{\text{in}} - \kappa_{\text{out}} > 2\varepsilon_f$).*

Theorem 3 (Quantitative Separability under Concept Shift). *If class c on client i undergoes concept shift toward semantic coordinate $j^\star = \pi_i(c) \neq c$, the in-class NTK match at $j^\star$ modifies the CAM entry $\mathbf{M}_{c,j^\star}^{(i)}$ relative to a normal reference client. Whenever the assumed amplification gap dominates statistical fluctuation, the row-wise ℓ_2 distance between the two empirical CAM rows admits a strictly positive lower bound with high probability: $\|\mathbf{M}_{c,:}^{(i)} - \mathbf{M}_{c,:}^{(0)}\|_2 \geq \Delta_c(T, \delta) > 0$.*

The full assumptions, lemmas, and complete proofs are provided in Appx. A.4. We further conduct controlled experiments to empirically validate all three theorems as presented in Appx. B.1.

3.2 FedCap: A Fine-Grained Utility-Driven CFL Framework

Following CAM construction, FedCap uses these utility profiles to group clients and guide cross-cluster knowledge sharing. The procedures are detailed below.

3.2.1 CAM-based Angular Distance

Each CAM row is normalized independently: $\mathbf{v}_c^{(i)} = \mathbf{M}_{c,:}^{(i)} / \|\mathbf{M}_{c,:}^{(i)}\|_2$, with zero rows for missing classes or negligible norms. We concatenate them in fixed class order as $\mathbf{m}_i = [\mathbf{v}_1^{(i)}; \ldots; \mathbf{v}_C^{(i)}] \in \mathbb{R}^{C^2}$. This reduces scale effects while preserving class-to-class interactions. The angular distance between nonzero client signatures is

$$d_{i,j} = \cos^{-1}\left(\frac{\langle \mathbf{m}_i, \mathbf{m}_j \rangle}{\|\mathbf{m}_i\|_2 \|\mathbf{m}_j\|_2}\right) \times \frac{180}{\pi}, \quad \forall i, j \in \{1, \ldots, N\}, \tag{6}$$

The off-diagonal distances are rank-normalized to form a symmetric dissimilarity matrix $\mathbf{D} \in [0,1]^{N \times N}$ with zero diagonal. Hierarchical clustering and silhouette scoring use this rank scale, rather than degrees. Numerical conventions and sparse transmission are described in Appx. B.2.

3.2.2 Conflict-Aware Hierarchical Clustering

Under severe non-IID conditions, clients that appear geometrically proximate may still be functionally incompatible, leading to *negative transfer*. To mitigate this, FedCap introduces an adaptive clustering mechanism governed by a novel CAM-guided fitness objective. Given a candidate clustering configuration $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ generated by Hierarchical Clustering with distance threshold α, we evaluate its quality using three components:

$$S(\mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} \frac{b(i) - a(i)}{\max a(i), b(i)}, \quad P_{size}(\mathcal{C}) = \frac{1}{K}\sum_{k=1}^{K} \exp\left(\frac{\max\left(0, \bar{s} - \gamma\sigma_s - |\mathcal{C}_k|\right)}{\tau}\right), \tag{7}$$

where $S(\mathcal{C})$ is the Silhouette Score [44]. For each client, $a(i)$ denotes the mean intra-cluster distance (compactness), and $b(i)$ denotes the mean distance to the nearest neighboring cluster (separation). For P_{size}, $\bar{s}$ is the mean cluster size, σ_s is the standard deviation of cluster sizes, and $\gamma \geq 0$ is a tolerance hyperparameter. Furthermore, we define a directed conflict score $\Phi(u \to v)$ that measures negative reference-class attribution from client u on classes present at client v:

$$\Phi(u \to v) = \sum_{k=1}^{C}\sum_{c=1}^{C} w_{u,k} \max(0, -\mathbf{S}_{u,k,c})\mathbf{O}_{v,c}, \tag{8}$$

where $w_{u,k}$ is the data proportion of class k on client u, $\mathbf{S}_{u,k,c}$ is the value from the row-wise normalized CAM indicating the specific impact of training source class k on target class c, and $\mathbf{O}_{v,c} \in \{0,1\}$ is the ownership mask indicating whether client v possesses class c. By truncating with $\max(0,-x)$, we isolate pure destructive interference, preventing it from being masked by positive complementary effects. Each client reports its class proportions and presence mask once with its CAM; $\mathbf{O}_{i,c} = \mathbb{I}[w_{i,c} > 0]$. The normalized conflict rate $R_{conf}(\mathcal{C})$ across all clusters is then calculated by averaging these directed conflicts over all valid intra-cluster pairs:

$$R_{conf}(\mathcal{C}) = \frac{1}{\sum_{k=1}^{K} |\mathcal{C}_k|(|\mathcal{C}_k|-1)} \sum_{k=1}^{K} \sum_{u \in \mathcal{C}_k} \sum_{\substack{v \in \mathcal{C}_k \\ v \neq u}} \Phi(u \to v). \tag{9}$$

With no non-self intra-cluster pairs, we set $R_{conf}(\mathcal{C}) = 0$. Zero conflict denotes no CAM-detected negative interaction, not guaranteed absence of negative transfer. The server sweeps thresholds on the rank-normalized dissimilarity matrix to generate candidate clusterings and selects α^* by the following objective:

$$\mathcal{L}_{cluster}(\alpha) = -S(\mathcal{C}^\alpha) + \lambda_1 R_{conf}(\mathcal{C}^\alpha) + \lambda_2 P_{size}(\mathcal{C}^\alpha), \tag{10}$$

where λ_1 and λ_2 are balancing hyperparameters. We select the lowest-objective admissible candidate, balancing separation, estimated conflict, and size. This is a finite search, not global partition optimization; see Appx. B.2.

3.2.3 Supply-Demand Guided Cross-Cluster Knowledge Sharing

After the warm-up phase, we further leverage the CAMs to orchestrate targeted knowledge transfer across clusters. The key insight is that each cluster simultaneously *supplies* beneficial gradient knowledge for certain classes while *demanding* complementary knowledge for others. FedCap uses CAMs and class proportions to select complementary clusters. Raw samples and per-sample gradients remain local, but server-visible metadata are not covered by a formal privacy guarantee.

Cluster-Level Knowledge Synergy. Using the row-normalized CAM $\mathbf{S}$, define client impact $a_{i,c} = \sum_{k=1}^{C} w_{i,k}\mathbf{S}_{i,k,c}$ and cluster impact $z_{p,c} = |\mathcal{C}_p|^{-1} \sum_{i \in \mathcal{C}_p} a_{i,c}$. The supply, demand, and directed supplier score are

$$s_{p,c} = [z_{p,c}]_+, \qquad d_{p,c} = [-z_{p,c}]_+, \qquad g_{p,q} = \sum_{c=1}^{C} d_{p,c} s_{q,c},$$

where $[x]_+ = \max(0,x)$. We select up to k positive-scoring suppliers $q \neq p$ with the highest $g_{p,q}$, forming $\mathcal{S}_p$. The supplier limit k differs from the cluster count K. The scores select suppliers; the selected encoders are averaged uniformly with $\alpha_{p,q} = 1/|\mathcal{S}_p|$:

$$\bar{\theta}^{(p)}_{syn} = \sum_{q \in \mathcal{S}_p} \alpha_{p,q} \theta^{(q)}, \tag{11}$$

Supplier sets and weights are fixed after warm-up; the target uses the latest cluster encoders $\theta^{(q)}$ each round. Empty supplier sets disable alignment. During local training, client $i \in \mathcal{C}_p$ initializes with its intra-cluster global model and explicitly absorbs this complementary remote knowledge via an alignment regularizer. The local objective for the model $W_i = \{\theta_i, \phi_i\}$ is formulated as:

$$\ell_i(W_i) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}\left[\mathcal{L}_{CE}(W_i;\mathbf{x},y)\right] + \frac{\mu}{2}\left\|\theta_i - \bar{\theta}^{(p)}_{syn}\right\|_2^2, \tag{12}$$

The coefficient μ controls encoder alignment. Both encoder and classifier are aggregated within each cluster; classifier heads are excluded from cross-cluster alignment.

Theorem 4 (Convergence of FedCap). Under standard smoothness, bounded-variance, and bounded-heterogeneity assumptions, the averaged iterate sequence generated by FedCap achieves an $\mathcal{O}(1/\sqrt{T})$ convergence rate to a stationary point neighborhood:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(W_t)\|_2^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}(\mu^2\Delta^2). \tag{13}$$

Here, the fixed synergy regularizer μ introduces a bounded perturbation term proportional to the maximum cross-cluster model divergence Δ. The detailed proof is deferred to Appx. A.6.

4 Experiments

4.1 Experimental Setup

Datasets and models We evaluate our approach on four popular image classification datasets: Fashion-MNIST [55], SVHN [42], CIFAR-10 [19], and CIFAR-100 [19]. For relatively simple tasks, we adopt a standard CNN [20] backbone, whereas ResNet-18 [13] is employed for CIFAR-100. Furthermore, we construct a lightweight public validation set by randomly sampling 2 instances per class to compute IRDS. Each experiment is conducted three times. Details are deferred to Appx. B.2.

Hyperparameter Settings. Without loss of generality, we adopt a set of default hyperparameters throughout all experiments. During the warm-up phase for IRDS estimation, each client concatenates a mini-batch of public validation data with a batch size of 5. In the clustering stage, the loss weights are set to $\lambda_1 = 1.0$ and $\lambda_2 = 0.01$, where the penalty term $P_{\text{size}}(\mathcal{C})$ is designed to only discourage the formation of extremely small clusters. In the penalty term $P_{\text{size}}(\mathcal{C})$, we set tolerance hyperparameter $\gamma = 1.0$. During training, we by default absorb the Top-$k = 1$ supply cluster to simulate a limited communication bandwidth scenario, and set the regularization coefficient to $\mu = 0.1$.

Data Distributions. To evaluate the robustness of our framework in complex heterogeneous scenarios, we design three non-IID data distributions to simulate label skew, feature skew, and concept shift:

- *Distribution 1 (Label Skew & Quantity Shift):* To simulate simultaneous label skew and quantity shift, we partition the data using two strategies: a 30% pathological label skew and the Latent Dirichlet Allocation (LDA) method [15], both under Dirichlet factor of $\alpha' = 1.0$ and 0.25.
- *Distribution 2 (Feature Skew & LDA):* Following the settings in FedRC [12] and FedDAG [43], we simulate feature skew using CIFAR-10-C [14], PACS and Office-Caltech-10 [60]. Each client is randomly assigned a specific corruption type or domain, introducing distinct feature variations. These samples are further partitioned under Dirichlet $\alpha' = 1.0$ to simulate combined with label skews.
- *Distribution 3 (Concept Shift):* Following [12, 43], We simulate concept shift by systematically modifying the labels for a subset of clients (30%). For instance, a ground-truth label y is mapped to $(C - y)$ or $(y + 1)\%C$, where C is the total number of classes.

Baselines. We compare FedCap against a comprehensive set of state-of-the-art baselines across four categories: 1) *Single-model FL methods:* FedAvg [39] and FedProx [22]; 2) *Personalized FL methods:* Per-FedAvg [7] and FedAMP [16]; 3) *Clustered FL methods:* including data-based approaches like PACFL [50], and gradient/loss-based approaches such as IFCA [11], CFL [46], FedSoft [45], FedRC [12], and FedDAG [43].

4.2 Experiment Results

Main Results. Tab. 1 and Fig. 5 report the performance under *Distribution 1*. The 30% label skew (Tab. 1) inherently creates distinct cluster boundaries, a scenario highly advantageous for clustered FL. FedCap achieves the highest accuracy across all tasks. Furthermore, the radar chart in Fig. 5 validates that FedCap successfully maintains its robust superiority even under standard continuous LDA distributions.

Tab. 2 highlights the robustness of FedCap under *Distributions 2 & 3*. In scenarios involving feature skew and concept shift, baselines relying on data subspace or gradients suffer from semantic misalignment and negative transfer. Conversely, FedCap successfully leverages its CAM-based conflict-avoidance clustering and supply-demand orchestration to extract genuine feature dependencies, outperforming the baselines.

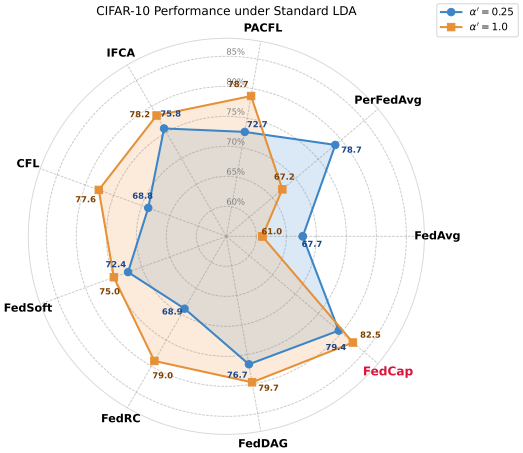

Figure 5: Comparison under standard LDA.

Table 1: Comparison of test accuracy under 30% Label Skew & Quantity Shift.

Algorithm	$\alpha' = 0.25$			$\alpha' = 1.0$		
	FMNIST	**CIFAR-10**	**CIFAR-100**	**FMNIST**	**CIFAR-10**	**CIFAR-100**
FedAvg	73.14 ± 0.38	52.73 ± 0.24	15.50 ± 1.28	78.05 ± 0.35	54.08 ± 0.70	20.18 ± 0.95
FedProx	69.05 ± 0.22	54.45 ± 0.22	15.70 ± 1.15	74.52 ± 0.40	54.56 ± 0.65	21.01 ± 0.88
PerFedAvg	89.42 ± 0.31	70.12 ± 0.28	27.65 ± 0.95	92.45 ± 0.28	85.54 ± 0.32	27.45 ± 0.75
FedAMP	87.92 ± 0.28	81.35 ± 0.35	28.52 ± 0.85	91.85 ± 0.25	81.25 ± 0.26	32.13 ± 0.65
PACFL	92.11 ± 0.15	83.67 ± 0.40	49.31 ± 0.52	93.12 ± 0.15	82.01 ± 0.18	42.18 ± 0.41
IFCA	91.31 ± 0.25	84.76 ± 0.30	48.76 ± 0.38	94.58 ± 0.12	78.25 ± 0.20	41.25 ± 0.34
CFL	82.54 ± 0.25	68.85 ± 0.16	38.45 ± 1.31	86.25 ± 0.22	77.62 ± 0.21	34.85 ± 0.88
FedSoft	84.15 ± 0.20	72.43 ± 0.22	42.15 ± 0.46	89.34 ± 0.18	76.95 ± 0.45	37.12 ± 0.35
FedRC	88.48 ± 0.32	76.92 ± 0.39	46.22 ± 0.35	89.85 ± 0.14	75.98 ± 0.22	39.45 ± 0.32
FedDAG	92.45 ± 0.12	89.65 ± 0.20	53.21 ± 0.58	94.62 ± 0.10	84.70 ± 0.17	43.81 ± 0.42
FedCap	**93.76** ± 0.42	**91.70** ± 0.18	**54.70** ± 0.45	**95.45** ± 0.26	**86.53** ± 0.14	**45.88** ± 0.35

Table 2: Performance comparison under combined feature skew (Left) and concept shift (Right).

Algorithm	Feature & Label Skew			Concept Shift		
	CIFAR-10-C	**PACS**	**Office-Caltech-10**	**CIFAR-10**	**FMNIST**	**SVHN**
FedAvg	31.47 ± 0.38	32.10 ± 0.31	41.87 ± 0.28	42.31 ± 0.34	41.57 ± 0.45	37.51 ± 0.36
PerFedAvg	59.88 ± 0.29	62.68 ± 0.34	62.55 ± 0.22	61.08 ± 0.25	74.05 ± 0.21	78.85 ± 0.25
PACFL	64.96 ± 0.25	67.85 ± 0.21	62.22 ± 0.17	58.53 ± 0.26	78.98 ± 0.38	80.17 ± 0.15
CFL	61.03 ± 0.18	60.72 ± 0.22	59.69 ± 0.21	62.32 ± 0.19	81.65 ± 0.26	79.54 ± 0.32
FedRC	60.74 ± 0.19	67.33 ± 0.15	64.66 ± 0.19	65.77 ± 0.35	79.18 ± 0.12	77.30 ± 0.31
IFCA	63.93 ± 0.42	63.82 ± 0.18	61.94 ± 0.33	67.56 ± 0.37	83.22 ± 0.18	82.67 ± 0.17
FedDAG	66.82 ± 0.35	68.79 ± 0.25	66.61 ± 0.24	68.05 ± 0.18	89.12 ± 0.15	84.62 ± 0.19
FedCap	**67.30** ± 0.22	**68.91** ± 0.19	**67.86** ± 0.27	**69.83** ± 0.21	**90.53** ± 0.14	**85.20** ± 0.26

Finally, Tab. 3 evaluates the server-side computational overhead for clustering 1000 clients. Unlike data-subspace or fusion-based methods that require computationally heavy matrix operations (e.g., SVD), our CAM-based distance metric is exceptionally efficient. This minimal computational footprint confirms that FedCap establishes a new SOTA in accuracy while remaining a highly lightweight solution for large-scale deployments.

Table 3: Server-side clustering time on 1000 clients (ResNet-18, CIFAR-100).

Method	**Time (s)** $\downarrow$
Gradient-based	0.287
Data (SVD subspace)	2.594
Fusion (Grad+Data)	3.070
CAM (Ours)	**0.098**

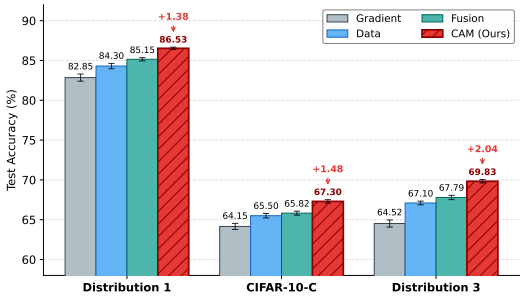

Figure 6: Ablation under *Distribution 1*.

Ablation. We conduct three sets of ablation studies to validate the core components of our framework. **1) Effectiveness of CAM:** Fig. 6 compares clustering metrics using our CAM versus purely gradient-based, data-based, and fused similarity proxies. The results clearly demonstrate CAM's superiority in capturing accurate client affinities. All similarity metrics are computed following the protocol in FadDAG[43], ensuring consistency in similarity estimation across different methods. **2) Conflict-Aware Clustering:** Tab. 4 evaluates the impact of our three clustering loss components. Relying solely on geometric compactness (Only $S(\mathcal{C}^\alpha)$) yields sub-optimal accuracy across all settings. Notably, removing the structural conflict penalty (W/o $R_{conf}(\mathcal{C}^\alpha)$) causes a distinct performance drop, especially under challenging Concept Shift scenarios. This highlights its critical role in filtering out gradient conflicts and preventing intra-cluster negative transfer. **3) Supply-Demand Synergy:** Results in Tab. 4 confirm the efficacy of our cross-cluster knowledge sharing module. Disabling the synergy target essentially reduces the framework to isolated clustered learning, leading to noticeable performance degradation. This validates that our supply-demand mechanism effectively mitigates cluster-level knowledge blind spots.

Table 4: Ablation studies on clustering losses and cross-cluster synergy. "w/o" denotes "without".

Algorithm	Label Skew & Quantity Shift		Feature Skew		Concept Shift	
	CIFAR-10	FMNIST	CIFAR-10-C	PACS	CIFAR-10	FMNIST
Only $S(\mathcal{C}^\alpha)$	84.79 ± 0.93	93.50 ± 0.65	65.15 ± 0.72	65.40 ± 0.81	67.80 ± 0.68	87.50 ± 0.55
W/o $P_{size}(\mathcal{C}^\alpha)$	84.29 ± 0.61	94.80 ± 0.52	66.20 ± 0.35	67.50 ± 0.32	68.10 ± 0.31	89.15 ± 0.28
W/o $R_{conf}(\mathcal{C}^\alpha)$	85.73 ± 0.21	94.60 ± 0.34	65.45 ± 0.44	65.88 ± 0.41	68.50 ± 0.52	88.20 ± 0.45
W/o synergy target	86.18 ± 0.38	94.20 ± 0.31	64.12 ± 0.38	66.75 ± 0.42	67.95 ± 0.27	88.45 ± 0.25
FedCap	**86.53** ± 0.14	**95.45** ± 0.26	**67.30** ± 0.22	**68.91** ± 0.19	**69.83** ± 0.21	**90.53** ± 0.14

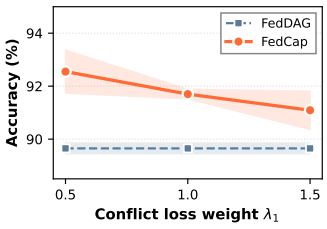

Figure 7: Impact of λ_1.

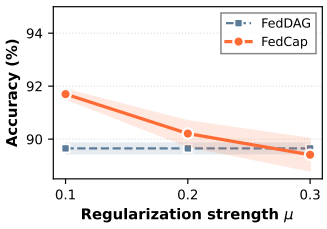

Figure 8: Impact of μ.

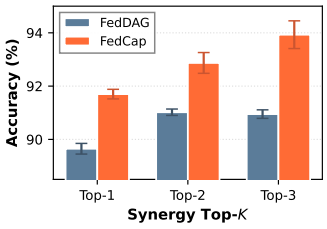

Figure 9: Impact of k.

Impact of hyperparameters. To investigate the sensitivity of FedCap to different hyperparameter configurations, we evaluate the performance of its key components under varying settings on CIFAR-10 with 30% label skew and a Dirichlet factor $\alpha = 0.25$. Specifically, we analyze the effects of the clustering conflict loss weight λ_1, the regularization strength μ for cross-cluster knowledge absorption, and the Top-K selection for cross-cluster synergy. As illustrated in Fig. 7, FedCap consistently outperforms state-of-the-art baseline across a wide range of λ_1 values, demonstrating its robustness. A smaller weight ($\lambda_1 = 0.5$) yields slightly better results by avoiding over-penalization. Regarding the regularization strength (Fig. 8), an excessively large μ restricts local updates, indicating that a mild constraint ($\mu = 0.1$) is optimal for knowledge absorption. Finally, as depicted in Fig. 9, integrating knowledge from broader sources (Top-3) significantly boosts FedCap's performance up to 93.93%, further widening the accuracy gap over FedDAG [43].

5 Related Work

Due to space constraints, we focus exclusively on the most closely related work here. A comprehensive discussion of broader literature is deferred to Appx. A.2.

Clustered FL (CFL) partitions clients into distinct groups based on specific affinity metrics to mitigate statistical heterogeneity, thereby training specialized models for different data distributions. The pioneering CFL framework [46, 61] evaluate client similarity based on model weights or gradients. However, these metrics inherently suffer from computational overhead and fail to group clients correctly due to the high dimensionality of data. Alternatively, data-centric approaches, notably PACFL [50], estimate similarity by extracting principal angles between client data subspaces. These methods risk privacy leakage, incur heavy matrix decomposition costs (e.g., SVD), and entirely ignore the clients' dynamic training capacities. To tackle diverse distribution shifts, recent frameworks have introduced robust soft clustering [45, 12], hybrid data-gradient fusion [43], and unsupervised descriptor extraction [8, 25]. In contrast, our approach explicitly quantifies fine-grained class-to-class interactions, effectively averting severe negative transfer in highly skewed scenarios.

Data Shapley [10] provides a principled, axiom-backed framework for equitable data valuation. However, traditional retraining-based methods and Monte Carlo approximations [17, 2, 23, 6] incur exponential computational costs, rendering them prohibitive for large-scale models. To break this bottleneck, In-Run Data Shapley (IRDS) [51] enables efficient attribution within a single training run by accumulating gradient dot-products as instantaneous marginal utilities. In FL, early Shapley frameworks [53, 48, 41] evaluate client contributions round-by-round, whereas Ripple Shapley [57] extends IRDS to capture temporal influence propagation via Jacobian tracking. Despite these advances, existing in-run FL attributions remain strictly confined to sample-level or coarse-grained

client valuation primarily for identifying low-quality data, failing to fully unleash the potential of value attribution. We leverage the IRDS philosophy to achieve client-side, class-to-class contribution attribution, explicitly extracting a class-wise interaction matrix to guide conflict-aware clustering and adaptive cross-cluster collaboration.

6 Conclusion and Limitations

In this paper, we introduced FedCap, a utility-driven Clustered FL framework that shifts the paradigm from similarity proxies to explicit value attribution. By leveraging the Class-interaction Attribution Matrix, FedCap effectively exposes hidden class conflicts, systematically mitigates negative transfer, and orchestrates cross-cluster synergy across diverse and complex distribution shifts. Despite its efficacy, our framework presents three main limitations:

1) Reliance on a public validation set: Computing the IRDS requires a class-balanced public validation set to serve as an unbiased global anchor. Although FedCap only necessitates a minimal data volume (e.g., 2-10 samples per class), acquiring such shared data remains a practical challenge in extreme few-shot or strictly privacy scenarios, which is a bottleneck of current attribution methods.

2) Fine-grained concept drift identification: While we have theoretically and empirically demonstrated CAM's capability to isolate concept-shifted clients from normal ones, distinguishing among diverse, concurrent concept drifts remains underexplored. Since different shifted clients may exhibit closely aligned cosine similarities, unleashing CAM's full potential for fine-grained concept drift identification and clustering represents a promising direction for our future work.

3) Metadata disclosure: Clients report CAMs, class proportions, and class-presence information to the server. These quantities can reveal distributional information despite raw samples remaining local. The evaluated method is unsuitable where such disclosure is prohibited and supplies no formal privacy guarantee.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China under Grant 2024YFC3307900; the National Natural Science Foundation of China under Grants 62376103, 62302184, 62436003, 62472332, and 62672205; the Major Science and Technology Project of Hubei Province under Grants 2025BAB011, 2024BAA008, and 2024BAB039; the Hubei Science and Technology Talent Service Project under Grant 2024DJC078; the Science and Technology Innovation Talent Program under grant 2023CB005; and Wuhan Dameng Database Co., Ltd.

References

[1] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[2] Mark Alexander Burgess and Archie C Chapman. Approximating the shapley value using stratified empirical bernstein sampling. In *IJCAI*, pages 73–81, 2021.

[3] Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

[4] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

[5] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021.

[6] Ian Covert, Chanwoo Kim, Su-In Lee, James Y. Zou, and Tatsunori B. Hashimoto. Stochastic amortization: A unified approach to accelerate feature and data attribution. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[7] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568, 2020.

[8] Dario Fenoglio, Mohan Li, Pietro Barbiero, Nicholas D Lane, Marc Langheinrich, and Martin Gjoreski. Flux: Efficient descriptor-driven clustered federated learning under arbitrary distribution shifts. *arXiv preprint arXiv:2511.22305*, 2025.

[9] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[10] Amirata Ghorbani and James Y. Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2242–2251, 2019.

[11] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[12] Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedrc: Tackling diverse distribution shifts challenge in federated learning by robust clustering. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pages 16910–16944, 2024.

[13] 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*, pages 770–778, 2016.

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[15] 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.

[16] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7865–7873, 2021.

[17] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd international conference on artificial intelligence and statistics*, pages 1167–1176, 2019.

[18] Myeongkyun Kang, Soopil Kim, Kyong Hwan Jin, Ehsan Adeli, Kilian M. Pohl, and Sanghyun Park. Fednn: Federated learning on concept drift data using weight and adaptive group normalizations. *Pattern Recognit.*, 149:110230, 2024.

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

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[21] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

[22] 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.

[23] Weida Li and Yaoliang Yu. Faster approximation of probabilistic and distributional values via least squares. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

[24] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[25] Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12820–12829. IEEE, 2024.

[26] Yichen Li, Qiyu Qin, Gaoyang Zhu, Wenchao Xu, Haozhao Wang, Yuhua Li, Rui Zhang, and Ruixuan Li. A systematic survey on federated sequential recommendation. *arXiv preprint arXiv:2504.05313*, 2025.

[27] Yichen Li, Yijing Shan, Yi Liu, Haozhao Wang, Cheng Wang, Yi Wang, and Ruixuan Li. Efficient knowledge transfer in federated recommendation for joint venture ecosystem. *Advances in Neural Information Processing Systems*, 38:124805–124826, 2026.

[28] Yichen Li, Yijing Shan, Yi Liu, Haozhao Wang, Wei Wang, Yi Wang, and Ruixuan Li. Personalized federated recommendation for cold-start users via adaptive knowledge fusion. In *Proceedings of the ACM on Web Conference 2025*, pages 2700–2709, 2025.

[29] Yichen Li, Hang Su, Huifa Li, Haolin Yang, Xinlin Zhuang, Haochen Xue, Haozhao Wang, and Imran Razzak. Fedcd: Towards consolidated distillation for heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23256–23264, 2026.

[30] Yichen Li, Haozhao Wang, Yining Qi, Wei Liu, and Ruixuan Li. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5489–5500, 2025.

[31] Yichen Li, Haozhao Wang, Wenchao Xu, Tianzhe Xiao, Hong Liu, Minzhu Tu, Yuying Wang, Xin Yang, Rui Zhang, Shui Yu, Song Guo, and Ruixuan Li. Unleashing the power of continual learning on non-centralized devices: A survey, 2024.

[32] Yichen Li, Xiuying Wang, Wenchao Xu, Haozhao Wang, Yining Qi, Jiahua Dong, and Ruixuan Li. Feature distillation is the better choice for model-heterogeneous federated learning. *Advances in Neural Information Processing Systems*, 38:104726–104744, 2026.

[33] Yichen Li, Yuying Wang, Jiahua Dong, Haozhao Wang, Yining Qi, Rui Zhang, and Ruixuan Li. Resource-constrained federated continual learning: What does matter? *Advances in Neural Information Processing Systems*, 38:79493–79518, 2026.

[34] Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. Fedssi: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *International Conference on Machine Learning*, pages 36151–36169. PMLR, 2025.

[35] Yichen Li, Wenchao Xu, Yining Qi, Haozhao Wang, Ruixuan Li, and Song Guo. Sr-fdil: Synergistic replay for federated domain-incremental learning. *IEEE Transactions on Parallel and Distributed Systems*, 35(11):1879–1890, 2024.

[36] Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In *European conference on computer vision*, pages 127–144. Springer, 2024.

[37] Alessandro Licciardi, Davide Leo, Eros Fanì, Barbara Caputo, and Marco Ciccone. Interaction-aware gaussian weighting for clustered federated learning. *arXiv preprint arXiv:2502.03340*, 2025.

[38] Qiqi Liu, Jiaqiang Li, Yuchen Liu, Yaochu Jin, Lingjuan Lyu, Xiaohu Wu, and Han Yu. Personalized federated learning under local supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4069–4079, 2025.

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

[40] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

[41] Lokesh Nagalapatti and Ramasuri Narayanam. Game of gradients: Mitigating irrelevant clients in federated learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, February 2-9, 2021*, pages 9046–9054, 2021.

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4, 2011.

[43] Anik Pramanik, Murat Kantarcioglu, Vincent Oria, and Shantanu Sharma. Feddag: Clustered federated learning via global data and gradient integration for heterogeneous environments. *arXiv preprint arXiv:2602.23504*, 2026.

[44] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[45] Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8124–8131, 2022.

[46] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[47] LS Shapley. 17. a value for n-person games. In *Contributions to the Theory of Games, Volume II*, pages 307–318. Princeton University Press, 2016.

[48] Qiheng Sun, Xiang Li, Jiayao Zhang, Li Xiong, Weiran Liu, Jinfei Liu, Zhan Qin, and Kui Ren. Shapleyfl: Robust federated learning based on shapley value. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 2096–2108, 2023.

[49] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

[50] Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10043–10052, 2023.

[51] Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

[52] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10165–10173, 2021.

[53] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In *Federated Learning - Privacy and Incentive*, pages 153–167. 2020.

[54] Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

[55] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[56] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

[57] Dewen Zeng, Wenlong Tian, Haozhao Wang, Jianfeng Lu, Weijun Xiao, and Zhiyong Xu. Ripple shapley: Data influence attribution in one federated training run. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 28085–28093, 2026.

[58] Jianqing Zhang, Yang Liu, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Jian Cao. Pfllib: A beginner-friendly and comprehensive personalized federated learning library and benchmark. *Journal of Machine Learning Research*, 26(50):1–10, 2025.

[59] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 26311–26329, 2022.

[60] Youshan Zhang and Brian D Davison. Impact of imagenet model selection on domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pages 173–182, 2020.

[61] Yuxin Zhang, Haoyu Chen, Zheng Lin, Zhe Chen, and Jin Zhao. Fedac: An adaptive clustered federated learning framework for heterogeneous data. *arXiv preprint arXiv:2403.16460*, 2024.

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

A Technical Appendices

A.1 Details of Distribution Shifts

In Federated Learning (FL), the dataset residing on client i can be denoted as $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{|\mathcal{D}_i|}$, which is drawn from a local joint probability distribution $\mathcal{P}_i(\mathbf{x}, y)$ over the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$. When data is Non-Identically and Independently Distributed (Non-IID) across the federated network, it implies that $\mathcal{P}_i(\mathbf{x}, y) \neq \mathcal{P}_j(\mathbf{x}, y)$ for different clients i and j. By factoring the joint distribution as $\mathcal{P}(\mathbf{x}, y) = \mathcal{P}(y|\mathbf{x})\mathcal{P}(\mathbf{x}) = \mathcal{P}(\mathbf{x}|y)\mathcal{P}(y)$, we can rigorously categorize the complex statistical heterogeneity into four distinct types of distribution shifts, as identified in [29, 18]:

• **Label Distribution Skew (Label Shift):** This shift occurs when the marginal distributions of labels vary across clients, i.e., $\mathcal{P}_i(y) \neq \mathcal{P}_j(y)$, even if the conditional distribution of features given a label remains the same, $\mathcal{P}_i(\mathbf{x}|y) = \mathcal{P}_j(\mathbf{x}|y)$. *Practical Example:* In a federated image classification task for animals, one client might predominantly capture images of dogs, while another client only has images of cats, resulting in non-identical class proportions [59].

• **Feature Distribution Skew (Feature Shift):** This shift happens when the marginal distributions of features differ across clients, i.e., $\mathcal{P}_i(\mathbf{x}) \neq \mathcal{P}_j(\mathbf{x})$, but the conditional distribution of labels given the features is shared, $\mathcal{P}_i(y|\mathbf{x}) = \mathcal{P}_j(y|\mathbf{x})$. Following the $Y \rightarrow X$ problem paradigm, this signifies a "same label, different features" scenario [40]. *Practical Example:* The same alphabet letter can be written in varying strokes and thicknesses due to different users' handwriting habits, or images of the same object may possess different lighting and contrast due to varying camera sensors [24].

• **Concept Shift:** Unlike label and feature skews where the decision boundary generally remains consistent, concept shift alters the underlying ground-truth mapping. It occurs when the conditional distributions of labels differ across clients, i.e., $\mathcal{P}_i(y|\mathbf{x}) \neq \mathcal{P}_j(y|\mathbf{x})$, even if $\mathcal{P}_i(\mathbf{x})$ and $\mathcal{P}_i(y)$ might be shared. In $X \rightarrow Y$ prediction tasks, this indicates a "same feature, different labels" scenario. *Practical Example:* In a sentiment analysis task, a specific word (e.g., "sick") might be labeled as a negative sentiment by older generations, but annotated as a highly positive sentiment (meaning "cool" or "excellent") by younger users. Consequently, forcing a single global model to fit these conflicting distributions leads to severe negative transfer.

• **Quantity Shift:** This form of heterogeneity simply refers to the severe unbalance in the volume of local data across the network, i.e., $|\mathcal{D}_i| \neq |\mathcal{D}_j|$. *Practical Example:* An enterprise-level hospital client might contribute hundreds of thousands of medical records to the federation, whereas a small community clinic might only possess a few hundred records [52, 27].

In real-world FL deployments, these four types of shifts rarely occur in isolation. Instead, they often manifest concurrently (e.g., simultaneous label skew and concept shift), creating highly complex optimization landscapes that bottleneck traditional single-global-model FL paradigms.

A.2 Additional Related Work

We provide a broader overview of the FL literature closely adjacent to our proposed framework.

Regularization and Collaborative Personalization. Ditto [21] regularizes personalized models toward a shared global model. pFedMe [49] uses Moreau envelopes to decouple personalized optimization from global learning in a bi-level objective. FedAMP [16] uses attentive model collaboration, while clustering methods such as FedAC [61], FedRC [12], FedSoft [45], and FedGWC [37] organize collaboration using different affinity signals. FedCap maintains cluster-level models and uses a warm-up class-interaction descriptor to select their collaboration partners.

Parameter Decoupling Strategies. FedPer [1] and FedRep [5] separate shared representations from client-specific prediction heads. FedRoD [4] and FedSimSup [38] use additional prediction or supervision mechanisms to balance shared and personalized learning. FedCap also separates encoder and classifier roles, but maintains separate cluster models and softly aligns encoders only to selected supplier clusters. Its heads are shared within clusters and excluded from cross-cluster alignment. This is a different sharing granularity from client-specific heads and a single shared representation.

Meta-Learning-Based Paradigms. Meta-learning adapts the Model-Agnostic Meta-Learning (MAML) framework for FL [35, 30] to find a well-generalized initialization for rapid local adaptation, as demonstrated in Per-FedAvg [7], pFedMe [49], and FedMeta [3]. Although effective under mild

heterogeneity, meta-learning algorithms typically require computation-heavy higher-order derivatives (e.g., Hessians) and struggle to converge when local task distributions diverge significantly. Our method circumvents these complex optimization landscapes by leveraging an efficient, single-run Shapley value to explicitly guide knowledge transfer without relying on second-order approximations.

A.3 Background of Data Shapley

The Shapley value was originally formulated in cooperative game theory [47] to address the problem of fairly distributing a total surplus or payout generated by a coalition of cooperating players. It provides a mathematically unique solution that assigns a value to each player strictly based on their average marginal contribution to every possible sub-coalition they could join.

Adaptation to Machine Learning. The concept was later adapted into the realm of machine learning, coining the term *Data Shapley* [10]. In this context, the "players" are the individual training data points (or in our case, the participating clients/classes), the "coalition" is the training dataset, and the "payout" (denoted by the utility function V) is the generalization performance of the machine learning model (e.g., accuracy or negative loss on a hold-out validation set). The Data Shapley value provides a principled framework to quantify the equitable value of each data source.

The widespread adoption of Data Shapley stems from the mathematical guarantee that it is the *only* valuation metric that simultaneously satisfies the following four desirable axioms. These axioms ensure that the attribution is fair, interpretable, and theoretically sound:

- **Null Player:** If a data point z_i provides exactly zero marginal change to the model's utility regardless of which subset it is added to (i.e., $V(S \cup \{z_i\}) = V(S)$ for all S), then its Shapley value is exactly zero ($\phi_i = 0$). This ensures that completely irrelevant or perfectly redundant data receives no credit.
- **Symmetry:** If two data points z_i and z_j always contribute exactly the same marginal utility to any arbitrary subset S that contains neither of them (i.e., $V(S \cup \{z_i\}) = V(S \cup \{z_j\})$), they must receive identical Shapley values ($\phi_i = \phi_j$). This axiom guarantees fairness by ensuring equal reward for equal contribution.
- **Linearity:** For any two utility functions V_1, V_2 and real scalars $\alpha, \beta \in \mathbb{R}$, the Shapley value of the combined utility is the linear combination of their individual Shapley values: $\phi_i(\alpha V_1 + \beta V_2) = \alpha\phi_i(V_1) + \beta\phi_i(V_2)$. This property is particularly crucial for *In-Run Data Shapley*, as it allows the overall attribution across a complex, multi-step training trajectory to be elegantly decomposed into a sum of instantaneous, step-wise marginal contributions.
- **Efficiency:** The sum of the Shapley values across all data points exactly equals the total utility of the model trained on the full dataset, minus the utility of a model trained on an empty set (e.g., a randomly initialized model before any training): $\sum_{i=1}^{N} \phi_i(V) = V(\mathcal{D}_{train}) - V(\emptyset)$. This allocates the full utility difference without unallocated utility; it is not a privacy property.

A.4 Theoretical Analysis of Class-interaction Attribution Matrix

In this appendix, we analyze CAM under explicit assumptions. The results address stability under label/quantity and bounded feature shifts, and raw-row separation under concept shift.

Preliminaries and Notation. Let C be the number of classes. For client i, let $\mathcal{D}_i$ denote its local training set and $\mathcal{D}_{\text{val}}$ the global class-balanced validation set broadcast by the server. For a local mini-batch trajectory of length T, the CAM of client i is defined as

$$\mathbf{M}^{(i)}_{c,c'} = \sum_{t=1}^{T} \sum_{(\mathbf{x},y)\in\mathcal{B}^{(t)}_{\text{train}},\, y=c} \Phi^{(t)}_{\mathbf{x}\to c'}, \quad \mathbf{M}^{(i)} \in \mathbb{R}^{C\times C}, \quad \mathbf{v}^{(i)}_c = \frac{\mathbf{M}^{(i)}_{c,:}}{\|\mathbf{M}^{(i)}_{c,:}\|_2}, \tag{14}$$

where $\Phi^{(t)}_{\mathbf{x}\to c'}$ is the accumulated attribution from a local training sample $\mathbf{x}$ to the validation class c' at step t, approximated in practice via the in-run gradient dot-product, and $\mathbf{v}^{(i)}_c$ is normalized row vector. We use these normalized row vectors as the geometric signatures for clustering. We further define the expected CAM:$\bar{\mathbf{M}}^{(i)} = \mathbb{E}[\mathbf{M}^{(i)}]$. The analysis is first conducted at the expectation level, and then lifted to the empirical CAM via concentration bounds.

A.4.1 Assumptions

Assumption 1 (NTK representation of gradient dot-products)**.** *For any two samples $z = (x, y)$ and $z' = (x', y')$, the full-parameter gradient dot-product of the cross-entropy loss satisfies*

$$\langle \nabla_\theta \ell(z;\theta), \nabla_\theta \ell(z';\theta) \rangle = (\mathbf{p}(x) - \mathbf{e}_y)^\top \mathbf{\Theta}(x, x')(\mathbf{p}(x') - \mathbf{e}_{y'}), \tag{15}$$

where $\mathbf{p}(x) \in \mathbb{R}^C$ is the softmax output, $\mathbf{e}_y$ is the one-hot label vector, and $\mathbf{\Theta}(x, x') := J_\theta(x) J_\theta(x')^\top \in \mathbb{R}^{C\times C}$ is the empirical neural tangent kernel (NTK), with $J_\theta(x)$ being the Jacobian of logits with respect to the full parameter vector θ.

Assumption 2 (Class-conditional NTK separation)**.** *Let $\gamma(x) \in \{1, \dots, C\}$ denote the* semantic *class of x, which may differ from its observed label under concept shift. There exist constants $\kappa_{\text{in}} > \kappa_{\text{out}} \geq 0$ such that, for all semantic classes a, b,*

$$\mathbb{E}[\mathbf{\Theta}(x, x') \mid \gamma(x) = a, \gamma(x') = b] = \mathbf{K}_{ab}, \text{ and } \|\mathbf{K}_{aa}\|_{\text{op}} \geq \kappa_{\text{in}},\ \|\mathbf{K}_{ab}\|_{\text{op}} \leq \kappa_{\text{out}},\ a \neq b. \tag{16}$$

Assumption 3 (Predictive-bias regimes induced by local label marginals)**.** *For client i, define the class-conditional average residual vector $\mathbf{r}_c^{(i)} = \mathbb{E}[\mathbf{p}(x) - \mathbf{e}_c \mid y = c,\ x \sim \mathcal{D}_i]$. We consider two regimes:*

(i) ***High-α (non-dominant) regime.*** *There exists $\eta_u > 0$ such that*

$$\left\| \mathbf{r}_c^{(i)} - \left(\frac{1}{C}\mathbf{1} - \mathbf{e}_c \right) \right\|_2 \leq \eta_u, \forall c. \tag{17}$$

(ii) ***Low-α (dominant-class) regime.*** *There exists a dominant class d_i and $\eta_d > 0$ such that*

$$\|\mathbf{r}_c^{(i)} - (\mathbf{e}_{d_i} - \mathbf{e}_c)\|_2 \leq \eta_d, \forall c \neq d_i, \text{ and } \|\mathbf{r}_{d_i}^{(i)}\|_2 \leq \eta_d. \tag{18}$$

Justification. Assumption 3 quantifies the well-documented phenomenon that local training under non-IID label marginals biases the classifier toward local dominant classes. The use of Dirichlet-distributed label proportions is standard in FL [15, 56, 24, 62]. The two regimes above explicitly cover both high-α clients without a dominant class and low-α clients with severe majority collapse.

Assumption 4 (Bounded feature skew perturbation)**.** *For a client i with feature skew, its class-conditional NTK satisfies*

$$\|\mathbf{K}_{ab}^{(i)} - \mathbf{K}_{ab}\|_{\text{op}} \leq \varepsilon_f, \forall a, b, \text{for some } \varepsilon_f \text{ obeying } 0 \leq \varepsilon_f < \kappa_{\text{in}} - \kappa_{\text{out}}. \tag{19}$$

Assumption 5 (Concept shift as semantic-label mismatch with NTK amplification)**.** *For a client i with concept shift, there exists a non-identity mapping $\pi_i : \{1, \dots, C\} \to \{1, \dots, C\}, \quad \pi_i(c) \neq c$ for at least one class c, such that local samples with observed label c have semantic class $\pi_i(c)$. Let $\mathbf{q}_{c'} = \mathbb{E}[\mathbf{p}(x') - \mathbf{e}_{c'} \mid x' \sim \mathcal{D}_{\text{val}},\ y' = c']$ denote the validation residual vector. We make two structural claims:*

(a) NTK-amplified magnitude at the misaligned coordinate. *For every shifted class c and $j^\star := \pi_i(c)$, there exists $m_{cs} > 0$ such that*

$$\left| \left(\mathbf{r}_c^{(i)}\right)^\top \mathbf{K}^{(i)}_{\pi_i(c),\pi_i(c)}\, \mathbf{q}_{j^\star} \right| \ \geq\ m_{cs} \ \propto\ \kappa_{\text{in}}. \tag{20}$$

(b) Bounded reference magnitude at the same coordinate. *For the normal reference client 0, since its samples with observed label c are semantically class c (no mismatch), the corresponding kernel block at the same coordinate $(c, j^\star)$ is the off-diagonal block $\mathbf{K}^{(0)}_{c,j^\star}$, hence*

$$\left| \left(\mathbf{r}_c^{(0)}\right)^\top \mathbf{K}^{(0)}_{c,c}\, \mathbf{q}_{j^\star} \right| \ \leq\ b_{cs} \ \propto\ \kappa_{\text{out}}. \tag{21}$$

The semantic kernel margin in Assumption 2 ensures $m_{cs} > b_{cs}$, i.e. $m_{cs} - b_{cs} \ \propto\ \kappa_{\text{in}} - \kappa_{\text{out}} \ >\ 0$.

Justification. Assumption 5 formalizes concept shift as a mismatch between the observed label and the underlying semantic class, consistent with the standard notion of concept drift in distribution-shift theory [9, 54, 40]. In the high-α predictive-bias regime of Assumption 3(i), both quantities may share the same sign, with the residual structure $(\mathbf{p} - \mathbf{e}_c)^\top (\mathbf{p} - \mathbf{e}_{c'})$ contributing a common factor; what distinguishes concept shift from normal behaviour is the kernel block: at $(c, j^\star)$ the shifted client invokes the in-class block $\mathbf{K}_{j^\star,j^\star}$, whereas the normal client invokes the off-diagonal block $\mathbf{K}_{c,j^\star}$.

Assumption 6 (Concentration of empirical CAM)**.** *For every client i, class pair (c, c'), and training step t, the one-step attribution contribution is almost surely bounded: $|\Phi^{(t)}_{\mathbf{x}\to c'}| \le B$. Moreover, the attribution sequence along the local trajectory forms either an independent sequence or a martingale difference sequence with bounded increments. Then for any $\delta \in (0,1)$, there exists a constant $K_B > 0$ such that*

$$\Pr\Big(\|\mathbf{M}^{(i)} - \bar{\mathbf{M}}^{(i)}\|_{\max} \le \xi_T(\delta)\Big) \ge 1-\delta, \quad \xi_T(\delta) := K_B\sqrt{T\log\frac{C^2}{\delta}}. \tag{22}$$

Justification. This assumption follows from standard bounded-difference concentration tools such as Hoeffding, Azuma-Hoeffding, or Bernstein inequalities. It allows us to lift expected CAM statements to high-probability guarantees for the empirical CAM.

A.4.2 Auxiliary Lemmas

Lemma 1 (Expected CAM decomposition)**.** *For each client i and any class pair (c, c'), the expected CAM entry admits the factorized form*

$$\bar{\mathbf{M}}^{(i)}_{c,c'} = N_c^{(i)} N_{c'}^{\mathrm{val}} \left(\mathbf{r}_c^{(i)}\right)^\top \mathbf{K}^{(i)}_{c,c'} \mathbf{q}_{c'}, \tag{23}$$

where $N_c^{(i)}$ is the number of local samples in class c on client i, and $N_{c'}^{\mathrm{val}}$ is the number of validation samples in class c'.

Proof. By the CAM definition in (14),

$$\bar{\mathbf{M}}^{(i)}_{c,c'} = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{(\mathbf{x},y)\in\mathcal{B}^{(t)}_{\mathrm{train}},\, y=c} \Phi^{(t)}_{\mathbf{x}\to c'}\right].$$

Using Assumption 1, each term can be written as

$$\Phi^{(t)}_{\mathbf{x}\to c'} = \sum_{(\mathbf{x}',y')\in\mathcal{B}^{(t)}_{\mathrm{val}},\, y'=c'} (\mathbf{p}(\mathbf{x}) - \mathbf{e}_c)^\top \mathbf{\Theta}(\mathbf{x},\mathbf{x}')(\mathbf{p}(\mathbf{x}') - \mathbf{e}_{c'}).$$

Taking class-conditional expectation and grouping by counts of local class-c samples and validation class-c' samples yields

$$\bar{\mathbf{M}}^{(i)}_{c,c'} = N_c^{(i)} N_{c'}^{\mathrm{val}} \mathbb{E}\big[(\mathbf{p}(\mathbf{x}) - \mathbf{e}_c)^\top \mathbf{\Theta}(\mathbf{x},\mathbf{x}')(\mathbf{p}(\mathbf{x}') - \mathbf{e}_{c'}) \mid y = c,\ y' = c'\big].$$

By the definitions of $\mathbf{r}_c^{(i)}$, $\mathbf{q}_{c'}$, and $\mathbf{K}^{(i)}_{c,c'}$, we obtain (23). □

Lemma 2 (Cosine stability under vector perturbation)**.** *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ be nonzero vectors. If*

$$\|\mathbf{a}-\mathbf{b}\|_2 \le \epsilon \qquad \textit{and} \qquad \min\{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2\} \ge m_0 > 0,$$

then

$$\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\rangle \ge 1 - \frac{2\epsilon^2}{m_0^2}. \tag{24}$$

Proof. By the triangle inequality,

$$\begin{aligned}\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2 &\le \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{a}\|_2} \right\|_2 + \left\| \frac{\mathbf{b}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2 \\ &= \frac{\|\mathbf{a}-\mathbf{b}\|_2}{\|\mathbf{a}\|_2} + \|\mathbf{b}\|_2 \left| \frac{1}{\|\mathbf{a}\|_2} - \frac{1}{\|\mathbf{b}\|_2} \right|.\end{aligned}$$

Since

$$\big|\|\mathbf{a}\|_2 - \|\mathbf{b}\|_2\big| \le \|\mathbf{a}-\mathbf{b}\|_2,$$

we further have

$$\|\mathbf{b}\|_2 \left| \frac{1}{\|\mathbf{a}\|_2} - \frac{1}{\|\mathbf{b}\|_2} \right| = \frac{\big|\|\mathbf{a}\|_2 - \|\mathbf{b}\|_2\big|}{\|\mathbf{a}\|_2} \le \frac{\|\mathbf{a}-\mathbf{b}\|_2}{\|\mathbf{a}\|_2}.$$

Therefore,

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2 \leq \frac{2\|\mathbf{a}-\mathbf{b}\|_2}{\|\mathbf{a}\|_2} \leq \frac{2\epsilon}{m_0}.$$

Using the identity

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2^2 = 2 - 2\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\rangle,$$

we obtain

$$2 - 2\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\rangle \leq \frac{4\epsilon^2}{m_0^2}.$$

Rearranging gives

$$\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\rangle \geq 1 - \frac{2\epsilon^2}{m_0^2},$$

which proves (24). □

Proof. By triangle inequality and norm continuity,

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2 \leq \frac{2\|\mathbf{a}-\mathbf{b}\|_2}{m_0} \leq \frac{2\epsilon}{m_0}, \text{using} \left\| \frac{\mathbf{a}}{\|\mathbf{a}\|_2} - \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\|_2^2 = 2 - 2\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_2}, \frac{\mathbf{b}}{\|\mathbf{b}\|_2} \right\rangle,$$

we obtain (24). □

A.4.3 Main Theorems

Theorem 1 (Structural isomorphism under quantity shift and label skew)**.** *Consider two clients i and j without concept shift. Assume they share the same predictive-bias regime in Assumption 3, i.e., either both are in the high-α regime, or both are in the low-α regime with the same dominant class. Suppose further that*

$$\sup_c \|\mathbf{r}_c^{(i)} - \mathbf{r}_c^{(j)}\|_2 \leq \varepsilon_r, \qquad \sup_{a,b} \|\mathbf{K}_{ab}^{(i)} - \mathbf{K}_{ab}^{(j)}\|_{\mathrm{op}} \leq \varepsilon_k. \tag{25}$$

Then for every class c locally present on both clients,

$$\left\| \frac{\bar{\mathbf{M}}_{c,:}^{(i)}}{\|\bar{\mathbf{M}}_{c,:}^{(i)}\|_2} - \frac{\bar{\mathbf{M}}_{c,:}^{(j)}}{\|\bar{\mathbf{M}}_{c,:}^{(j)}\|_2} \right\|_2 \leq \frac{L_c(\varepsilon_r + \varepsilon_k)}{m_c}, \tag{26}$$

where

$$m_c := \min\{\|\bar{\mathbf{M}}_{c,:}^{(i)}\|_2, \|\bar{\mathbf{M}}_{c,:}^{(j)}\|_2\} > 0,$$

and $L_c > 0$ depends only on upper bounds of $\|\mathbf{r}_c^{(i)}\|_2$, $\|\mathbf{r}_c^{(j)}\|_2$, $\|\mathbf{q}_{c'}\|_2$, and class counts.

Consequently,

$$\left\langle \bar{\mathbf{v}}_c^{(i)}, \bar{\mathbf{v}}_c^{(j)} \right\rangle \geq 1 - \frac{2L_c(\varepsilon_r + \varepsilon_k)}{m_c}. \tag{27}$$

Moreover, under Assumption 6, with probability at least $1-\delta$,

$$\left\langle \mathbf{v}_c^{(i)}, \mathbf{v}_c^{(j)} \right\rangle \geq 1 - \frac{2L_c(\varepsilon_r + \varepsilon_k)}{m_c} - \frac{4\sqrt{C}\,\xi_T(\delta)}{m_c}. \tag{28}$$

In particular, quantity shift alone only rescales $\bar{\mathbf{M}}_{c,:}^{(i)}$ and therefore disappears after row normalization.

Proof. By Lemma 1,

$$\bar{\mathbf{M}}_{c,c'}^{(i)} = N_c^{(i)} N_{c'}^{\mathrm{val}} (\mathbf{r}_c^{(i)})^\top \mathbf{K}_{c,c'}^{(i)} \mathbf{q}_{c'},$$

and analogously for client j. For a fixed c',

$$\begin{aligned}
&\left| \frac{\bar{\mathbf{M}}^{(i)}_{c,c'}}{N^{(i)}_c N^{\mathrm{val}}_{c'}} - \frac{\bar{\mathbf{M}}^{(j)}_{c,c'}}{N^{(j)}_c N^{\mathrm{val}}_{c'}} \right| \\
&= \left| (\mathbf{r}^{(i)}_c)^\top \mathbf{K}^{(i)}_{c,c'} \mathbf{q}_{c'} - (\mathbf{r}^{(j)}_c)^\top \mathbf{K}^{(j)}_{c,c'} \mathbf{q}_{c'} \right| \\
&\le \left| (\mathbf{r}^{(i)}_c - \mathbf{r}^{(j)}_c)^\top \mathbf{K}^{(i)}_{c,c'} \mathbf{q}_{c'} \right| + \left| (\mathbf{r}^{(j)}_c)^\top (\mathbf{K}^{(i)}_{c,c'} - \mathbf{K}^{(j)}_{c,c'}) \mathbf{q}_{c'} \right| \\
&\le \|\mathbf{r}^{(i)}_c - \mathbf{r}^{(j)}_c\|_2 \|\mathbf{K}^{(i)}_{c,c'}\|_{\mathrm{op}} \|\mathbf{q}_{c'}\|_2 + \|\mathbf{r}^{(j)}_c\|_2 \|\mathbf{K}^{(i)}_{c,c'} - \mathbf{K}^{(j)}_{c,c'}\|_{\mathrm{op}} \|\mathbf{q}_{c'}\|_2 \\
&\le L_{c,c'}(\varepsilon_r + \varepsilon_k),
\end{aligned}$$

for some constant $L_{c,c'} > 0$. Summing over all $c' = 1, \dots, C$ gives

$$\left\| \frac{\bar{\mathbf{M}}^{(i)}_{c,:}}{N^{(i)}_c} - \frac{\bar{\mathbf{M}}^{(j)}_{c,:}}{N^{(j)}_c} \right\|_2 \le L_c(\varepsilon_r + \varepsilon_k).$$

Quantity shift only changes the positive scalar $N^{(i)}_c$, which vanishes after row normalization. Therefore,

$$\left\| \frac{\bar{\mathbf{M}}^{(i)}_{c,:}}{\|\bar{\mathbf{M}}^{(i)}_{c,:}\|_2} - \frac{\bar{\mathbf{M}}^{(j)}_{c,:}}{\|\bar{\mathbf{M}}^{(j)}_{c,:}\|_2} \right\|_2 \le \frac{L_c(\varepsilon_r + \varepsilon_k)}{m_c},$$

which proves (26). Applying Lemma 2 yields (27).

To lift the result to empirical CAM, note that by Assumption 6,

$$\|\mathbf{M}^{(i)} - \bar{\mathbf{M}}^{(i)}\|_2 \le \sqrt{C}\, \xi_T(\delta),$$

row-wise, with probability at least $1 - \delta$, and similarly for client j. Another application of Lemma 2 gives (28). □

Theorem 2 (Topological robustness under feature skew)**.** *Consider a client i with feature skew but without concept shift, and a reference normal client 0 that shares the same predictive-bias regime and label marginal type. Under Assumption 4, for every locally present class c,*

$$\left\| \bar{\mathbf{v}}^{(i)}_c - \bar{\mathbf{v}}^{(0)}_c \right\|_2 \le \frac{L'_c\, \varepsilon_f}{m'_c}, \tag{29}$$

where

$$m'_c := \min\{\|\bar{\mathbf{M}}^{(i)}_{c,:}\|_2, \|\bar{\mathbf{M}}^{(0)}_{c,:}\|_2\} > 0.$$

Consequently,

$$\left\langle \bar{\mathbf{v}}^{(i)}_c, \bar{\mathbf{v}}^{(0)}_c \right\rangle \ge 1 - \frac{2L'_c \varepsilon_f}{m'_c}. \tag{30}$$

Furthermore, if

$$\kappa_{\mathrm{in}} - \kappa_{\mathrm{out}} > 2\varepsilon_f, \tag{31}$$

then the diagonal-dominance topology is preserved, i.e., there exists $\Delta_c > 0$ such that

$$\bar{\mathbf{M}}^{(i)}_{c,c} - \max_{c' \ne c} \bar{\mathbf{M}}^{(i)}_{c,c'} \ge \Delta_c > 0. \tag{32}$$

Under Assumption 6, with probability at least $1 - \delta$,

$$\left\langle \mathbf{v}^{(i)}_c, \mathbf{v}^{(0)}_c \right\rangle \ge 1 - \frac{2L'_c \varepsilon_f}{m'_c} - \frac{4\sqrt{C}\, \xi_T(\delta)}{m'_c}. \tag{33}$$

Proof. By Lemma 1, the difference between the expected CAM rows is entirely controlled by the perturbation of the kernel:

$$\bar{\mathbf{M}}^{(i)}_{c,c'} - \bar{\mathbf{M}}^{(0)}_{c,c'} = N^{(i)}_c N^{\mathrm{val}}_{c'} (\mathbf{r}^{(i)}_c)^\top (\mathbf{K}^{(i)}_{c,c'} - \mathbf{K}^{(0)}_{c,c'}) \mathbf{q}_{c'}.$$

Since the predictive-bias regime is shared and there is no concept shift, the residual vectors remain aligned, and Assumption 4 implies

$$|\bar{\mathbf{M}}^{(i)}_{c,c'} - \bar{\mathbf{M}}^{(0)}_{c,c'}| \le L'_{c,c'}\varepsilon_f.$$

Summing over c' gives

$$\|\bar{\mathbf{M}}^{(i)}_{c,:} - \bar{\mathbf{M}}^{(0)}_{c,:}\|_2 \le L'_c\varepsilon_f.$$

Equation (29) follows after normalization, and (30) is again obtained by Lemma 2.

To prove topology preservation, note that without feature skew, the semantic margin satisfies

$$\bar{\mathbf{M}}^{(0)}_{c,c} - \bar{\mathbf{M}}^{(0)}_{c,c'} \gtrsim \kappa_{\text{in}} - \kappa_{\text{out}}, \qquad \forall c' \neq c.$$

A perturbation bounded by ε_f on both the diagonal and off-diagonal entries can alter this gap by at most $2\varepsilon_f$. Therefore, (31) guarantees the remaining gap is still positive, proving (32). The empirical statement (33) follows from Assumption 6 exactly as in Theorem 1. □

Theorem 3 (Quantitative separability under concept shift)**.** *Consider a client i for which class c is affected by concept shift, with semantic class $\pi_i(c) = j^\star \neq c$, and let client 0 be a normal reference client. Assume there exist constants $A_c > B_c > 0$ such that*

$$\left|\, \bar{\mathbf{M}}^{(i)}_{c,j^\star} \,\right| \;\ge\; A_c, \text{and,} \quad \left|\, \bar{\mathbf{M}}^{(0)}_{c,j^\star} \,\right| \;\le\; B_c, \tag{34}$$

and that $\bar{\mathbf{M}}^{(i)}_{c,j^\star}$ and $\bar{\mathbf{M}}^{(0)}_{c,j^\star}$ have the same sign. Then the population-level row-wise ℓ_2 distance satisfies

$$\left\|\bar{\mathbf{M}}^{(i)}_{c,:} - \bar{\mathbf{M}}^{(0)}_{c,:}\right\|_2 \;\ge\; A_c - B_c \;>\; 0. \tag{35}$$

Moreover, under Assumption 6, if

$$A_c - B_c \;>\; 2\,\xi_T(\delta), \tag{36}$$

then with probability at least $1-\delta$,

$$\left\|\mathbf{M}^{(i)}_{c,:} - \mathbf{M}^{(0)}_{c,:}\right\|_2 \;\ge\; A_c - B_c - 2\,\xi_T(\delta) \;>\; 0. \tag{37}$$

Proof. **Population-level bound.** Since ℓ_2 distance majorises any single-coordinate gap,

$$\left\|\bar{\mathbf{M}}^{(i)}_{c,:} - \bar{\mathbf{M}}^{(0)}_{c,:}\right\|_2 \;\ge\; \left|\, \bar{\mathbf{M}}^{(i)}_{c,j^\star} - \bar{\mathbf{M}}^{(0)}_{c,j^\star} \,\right|.$$

Because $\bar{\mathbf{M}}^{(i)}_{c,j^\star}$ and $\bar{\mathbf{M}}^{(0)}_{c,j^\star}$ share the same sign, their difference satisfies

$$\left|\, \bar{\mathbf{M}}^{(i)}_{c,j^\star} - \bar{\mathbf{M}}^{(0)}_{c,j^\star} \,\right| \;=\; \left|\, \bar{\mathbf{M}}^{(i)}_{c,j^\star} \,\right| - \left|\, \bar{\mathbf{M}}^{(0)}_{c,j^\star} \,\right| \;\ge\; A_c - B_c,$$

where the first equality uses sign agreement and the inequality uses (34). This establishes (35).

Empirical bound. Define the deviations $\Delta_i := \mathbf{M}^{(i)}_{c,:} - \bar{\mathbf{M}}^{(i)}_{c,:}$ and $\Delta_0 := \mathbf{M}^{(0)}_{c,:} - \bar{\mathbf{M}}^{(0)}_{c,:}$. By Assumption 6, with probability at least $1-\delta$,

$$\|\Delta_i\|_\infty \le \xi_T(\delta), \qquad \|\Delta_0\|_\infty \le \xi_T(\delta).$$

Restricting attention to the $j^\star$-th coordinate and applying the reverse triangle inequality,

$$\left|\mathbf{M}^{(i)}_{c,j^\star} - \mathbf{M}^{(0)}_{c,j^\star}\right| \;\ge\; \left|\bar{\mathbf{M}}^{(i)}_{c,j^\star} - \bar{\mathbf{M}}^{(0)}_{c,j^\star}\right| - |\Delta_i|_{j^\star} - |\Delta_0|_{j^\star} \;\ge\; (A_c - B_c) - 2\xi_T(\delta).$$

Hence

$$\left\|\mathbf{M}^{(i)}_{c,:} - \mathbf{M}^{(0)}_{c,:}\right\|_2 \;\ge\; \left|\mathbf{M}^{(i)}_{c,j^\star} - \mathbf{M}^{(0)}_{c,j^\star}\right| \;\ge\; (A_c - B_c) - 2\xi_T(\delta).$$

Under (36) this lower bound is strictly positive, yielding (37). □

A.4.4 Discussion of the Assumptions

Remark 1 (Why the theory covers both high-α and low-α clients)**.** *Assumption 3 explicitly distinguishes the high-α regime, where no dominant class exists and the local predictor remains approximately uniform, from the low-α regime, where a dominant class induces a strong predictive bias. Therefore, both mild and extreme label skew are included in the same framework.*

Remark 2 (What makes concept shift unique)**.** *Quantity shift only rescales rows, label skew changes the residual pattern but preserves the semantic-label alignment, and bounded feature skew only perturbs the NTK geometry continuously. Concept shift changes the semantic-label correspondence. Under the amplification assumptions of Theorem 3, this produces a raw-entry magnitude gap relative to the reference client. The result does not require a sign reversal, nor does it establish a negative cosine margin after row normalization.*

A.5 FedCap Algorithmic Framework

To comprehensively illustrate the operational workflow of our proposed FedCap, we present the overall algorithmic framework in **Algorithm 1**.

Algorithm 1 Overall FedCap Framework

Input: Number of clients N, Public validation set $\mathcal{D}_{val}$, Warm-up rounds T_w, Total communication rounds T.
Output: Cluster-specific models $\{\bar{w}_T^{(k)}\}_{k=1}^K$

Phase I: Warm-up & Class-interaction Attribution
for round $t = 1, \ldots, T_w$ **do**
  for each client $i \in \{1, \ldots, N\}$ **in parallel do**
    Perform local SGD on $\mathcal{D}_i$; use $\mathcal{D}_{val}$ only for reference gradients
    Accumulate CAM $\mathbf{M}^{(i)}$ via training–validation ghost dot-products
  end for
end for
Clients send CAMs, class proportions w_i, presence masks $\mathbf{O}_i$, and data sizes $|\mathcal{D}_i|$ to the server once
Phase II: Clustering & Supply-demand Graph Construction
Server row-normalizes CAMs, flattens them, and rank-normalizes angular distances to form $\mathbf{D}$
Server selects $\mathcal{C}^*$ by the conflict-aware objective over admissible candidate partitions
Server computes supply s_p, demand d_p, and positive-score supplier sets $\mathcal{S}_p$ for each cluster p
Fix $\alpha_{p,q} = 1/|\mathcal{S}_p|$ for selected suppliers; disable alignment if $\mathcal{S}_p$ is empty
Phase III: Formal Training & Synergy
for round $t = T_w + 1, \ldots, T$ **do**
  Server forms synergy targets from current supplier models and selects participating clients
  for each participating client $i \in \mathcal{C}_p$ **in parallel do**
    Initialize encoder and classifier from the client's own cluster model $\bar{w}_{t-1}^{(p)}$
    Receive $\bar{\theta}_{syn,t-1}^{(p)}$ as the encoder-alignment target when available
    Update local model via SGD with proximal synergy regularization
  end for
  Server aggregates local updates to form new intra-cluster models $\bar{w}_t^{(k)}$ via FedAvg.
end for

A.6 Convergence Analysis of FedCap

In this appendix, we provide a theoretical analysis showing that the proposed synergy-guided federated optimization in FedCap converges stably to a neighborhood of the stationary point, preserving the standard convergence behavior of proximal federated learning with a bounded cross-cluster bias.

Preliminaries and Notation. Let θ denote the encoder parameters and let $F(\theta) := \frac{1}{N}\sum_{i=1}^N F_i(\theta)$ be the global encoder objective, where $F_i(\theta)$ is the local loss of client i. At communication round t, each cluster p has a clean intra-cluster aggregated encoder $\bar{\theta}_t^{(p)}$. Based on the normalized supply-demand scores, the server constructs a synergy target for cluster p by computing a weighted average of the global models from its assigned top-K supplier clusters $\mathcal{S}_p$:

$$\bar{\theta}_{syn,t}^{(p)} = \sum_{q \in \mathcal{S}_p} \alpha_{p,q} \bar{\theta}_t^{(q)}, \tag{38}$$

where $\alpha_{p,q} = 1/|\mathcal{S}_p|$ for nonempty supplier sets. If $\mathcal{S}_p$ is empty, the alignment regularizer is omitted and we set the cross-cluster perturbation below to zero by convention. For a client i in cluster p, the local encoder objective augmented with the synergy regularizer is

$$F_i^{\text{prox}}(\theta) := F_i(\theta) + \frac{\mu}{2}\|\theta - \bar{\theta}_{syn,t}^{(p)}\|_2^2. \tag{39}$$

We define the cross-cluster perturbation term as the discrepancy between the remote weighted synergy target and the local cluster's current global model:

$$\Delta_t^{(p)} := \bar{\theta}_{syn,t}^{(p)} - \bar{\theta}_t^{(p)} = \sum_{q \in \mathcal{S}_p} \alpha_{p,q}\big(\bar{\theta}_t^{(q)} - \bar{\theta}_t^{(p)}\big). \tag{40}$$

The analysis below shows that FedCap is equivalent to a proximal federated method with a bounded perturbation center.

Assumption 7 (Smooth local objectives)**.** *For every client i, the local objective $F_i(\theta)$ is L-smooth, i.e., for all θ, θ',*

$$F_i(\theta') \le F_i(\theta) + \langle \nabla F_i(\theta), \theta' - \theta \rangle + \frac{L}{2}\|\theta' - \theta\|_2^2. \tag{41}$$

Justification. Assumption 7 is the standard smoothness condition used in non-convex stochastic optimization and federated optimization analyses.

Assumption 8 (Unbiased stochastic gradients with bounded variance)**.** *For every client i, let $g_i(\theta;\xi)$ be the stochastic gradient computed on a mini-batch ξ. Then $\mathbb{E}_\xi[g_i(\theta;\xi)] = \nabla F_i(\theta)$, and there exists $\sigma^2 > 0$ such that*

$$\mathbb{E}_\xi \|g_i(\theta;\xi) - \nabla F_i(\theta)\|_2^2 \le \sigma^2. \tag{42}$$

Justification. Assumption 8 is the classical bounded-variance assumption ensuring that mini-batch sampling noise does not dominate the descent dynamics.

Assumption 9 (Bounded client heterogeneity)**.** *There exists $\zeta^2 > 0$ such that for all θ,*

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(\theta) - \nabla F(\theta)\|_2^2 \le \zeta^2. \tag{43}$$

Justification. Assumption 9 is standard in federated optimization to quantify statistical heterogeneity across clients, commonly used in FedAvg [39] and FedProx [22] analyses.

Assumption 10 (Bounded cross-cluster divergence)**.** *For every round t and cluster p, the magnitude of the cross-cluster perturbation defined in* (40) *is bounded:*

$$\|\Delta_t^{(p)}\|_2 \le \Delta, \tag{44}$$

for some constant $\Delta > 0$.

Justification. Assumption 10 is an explicit bounded-divergence condition, not an automatic property of unconstrained neural-network training. For nonempty supplier sets, convex weights imply $\|\Delta_t^{(p)}\|_2 \le \sum_q \alpha_{p,q}\|\bar\theta_t^{(q)} - \bar\theta_t^{(p)}\|_2$. Thus a uniform bound on pairwise cluster-model distances is sufficient.

Lemma 3 (FedCap as proximal federated optimization with bounded bias)**.** *For a client i in cluster p, the local objective in FedCap can be rewritten as*

$$F_i^{\mathrm{prox}}(\theta) = F_i(\theta) + \frac{\mu}{2}\|\theta - \bar\theta_t^{(p)}\|_2^2 + \mu\left\langle \theta - \bar\theta_t^{(p)}, -\Delta_t^{(p)} \right\rangle + \frac{\mu}{2}\|\Delta_t^{(p)}\|_2^2. \tag{45}$$

Consequently, FedCap optimizes a standard proximal objective centered at its own cluster model $\bar\theta_t^{(p)}$, plus a gradient bias proportional to $\mu\Delta_t^{(p)}$.

Proof. By definition, $\bar\theta_{syn,t}^{(p)} = \bar\theta_t^{(p)} + \Delta_t^{(p)}$. Substituting this into (39), we obtain $F_i^{\mathrm{prox}}(\theta) = F_i(\theta) + \frac{\mu}{2}\|\theta - \bar\theta_t^{(p)} - \Delta_t^{(p)}\|_2^2$. Expanding the squared norm gives $\|\theta - \bar\theta_t^{(p)} - \Delta_t^{(p)}\|_2^2 = \|\theta - \bar\theta_t^{(p)}\|_2^2 - 2\langle \theta - \bar\theta_t^{(p)}, \Delta_t^{(p)}\rangle + \|\Delta_t^{(p)}\|_2^2$, which proves (45). By Assumption 10, the gradient drift introduced by the third term is bounded by $\mu\Delta$. □

Theorem 4 (Convergence of FedCap)**.** *Consider the FedCap encoder optimization under Assumptions 7–10. Let $\{\theta_t\}_{t=0}^{T-1}$ be the sequence of global encoder iterates generated by FedCap with a constant stepsize η and a fixed regularization weight μ. Then the average expected squared gradient norm satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta_t)\|_2^2 \le \frac{F(\theta_0) - F^\star}{c_1 \eta T} + \frac{c_2 \eta(\sigma^2 + \zeta^2)}{c_1} + \frac{c_3}{c_1}\mu^2\Delta^2, \tag{46}$$

where $c_1, c_2, c_3 > 0$ are absolute constants, and $F^\star$ is the global minimum of F. Furthermore, by setting the constant stepsize appropriately as $\eta = \frac{1}{\sqrt{T}}$, FedCap achieves the following convergence rate:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \frac{A}{\sqrt{T}} + B\mu^2\Delta^2, \tag{47}$$

where $A = \frac{F(\theta_0)-F^\star}{c_1} + \frac{c_2(\sigma^2+\zeta^2)}{c_1}$ and $B = \frac{c_3}{c_1}$. That is, FedCap achieves an $\mathcal{O}(1/\sqrt{T})$ convergence rate to a stationary point neighborhood, bounded by the synergy bias $\mathcal{O}(\mu^2\Delta^2)$.

Proof. By Assumption 7, the objective function is L-smooth. For one global update step from θ_t to θ_{t+1}, we have

$$F(\theta_{t+1}) \leq F(\theta_t) + \langle \nabla F(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2. \tag{48}$$

Taking the conditional expectation on both sides and using the standard decomposition of stochastic local updates into the true global gradient, stochastic noise, client-drift error, and the gradient bias from Lemma 3, we obtain the following one-step descent lemma with the constant stepsize η:

$$\mathbb{E}[F(\theta_{t+1})] \leq \mathbb{E}[F(\theta_t)] - c_1\eta\,\mathbb{E}\|\nabla F(\theta_t)\|_2^2 + c_2\eta^2(\sigma^2+\zeta^2) + c_3\eta\mu^2\|\Delta_t^{(p)}\|_2^2, \tag{49}$$

for some positive constants c_1, c_2, c_3 that depend on the smoothness constant L and the local update steps. Under Assumption 10, the synergy bias is uniformly bounded by $\|\Delta_t^{(p)}\|_2^2 \leq \Delta^2$. Substituting this bound into (49) and rearranging the terms to isolate the expected squared gradient norm, we get:

$$c_1\eta\,\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \mathbb{E}[F(\theta_t)] - \mathbb{E}[F(\theta_{t+1})] + c_2\eta^2(\sigma^2+\zeta^2) + c_3\eta\mu^2\Delta^2. \tag{50}$$

Summing the above inequality over $t = 0, \ldots, T-1$ allows the objective function terms to telescope:

$$c_1\eta\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq F(\theta_0) - \mathbb{E}[F(\theta_T)] + c_2\eta^2T(\sigma^2+\zeta^2) + c_3\eta T\mu^2\Delta^2. \tag{51}$$

Since the global minimum $F^\star$ lower bounds $\mathbb{E}[F(\theta_T)]$ (i.e., $-\mathbb{E}[F(\theta_T)] \leq -F^\star$), we can replace the telescoping sum bound with $F(\theta_0) - F^\star$. Dividing both sides by $c_1\eta T$ directly yields the general bound for a constant stepsize:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \frac{F(\theta_0)-F^\star}{c_1\eta T} + \frac{c_2\eta}{c_1}(\sigma^2+\zeta^2) + \frac{c_3}{c_1}\mu^2\Delta^2. \tag{52}$$

To minimize the dependency on T for the first two terms, we specify the constant stepsize as $\eta = \frac{1}{\sqrt{T}}$. Substituting this into the inequality, the first term becomes $\frac{F(\theta_0)-F^\star}{c_1\sqrt{T}}$ and the second term becomes $\frac{c_2(\sigma^2+\zeta^2)}{c_1\sqrt{T}}$. Factoring out $1/\sqrt{T}$ provides:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\theta_t)\|_2^2 \leq \frac{A}{\sqrt{T}} + B\mu^2\Delta^2, \tag{53}$$

where $A = \frac{F(\theta_0)-F^\star+c_2(\sigma^2+\zeta^2)}{c_1}$ and $B = \frac{c_3}{c_1}$. This completes the proof, demonstrating that FedCap with a constant stepsize reliably converges to a stationary point neighborhood defined by the structural synergy bias. □

B Additional Experiments

B.1 Empirical Validation of CAM Properties

To validate the theoretical properties of the Class-interaction Attribution Matrix (CAM) established in Theorems 1–3, we conduct controlled experiments using SimpleCNN. We evaluate label/quantity shifts on CIFAR-10 (C=10, 100 clients) and feature skew on PACS (C=7, 50 clients). After a warm-up of T=10 local epochs, we extract the CAM $\mathbf{M}^{(i)}$ and compute the **row-wise cosine similarity** averaged over shared classes.

B.1.1 Validation of Theorem 1: Label/Quantity Shift Isomorphism

Setup. We partition 100 clients into four 25-client groups:

- **G1 (Balanced):** Dirichlet(α=1.0) across all 10 classes (weak predictive bias).
- **G2 (Dom-3):** Dirichlet(α=0.25) on $\{3, 6, 9\}$ only, with class 3 forced as the global dominant class for all clients.
- **G3 (Mixed-Dom):** Same as G2, but the dominant class varies across clients (forming dom=3, dom=6, dom=9 sub-clusters).
- **G4 (Quantity Shift):** Same distribution as G1, but local dataset sizes are randomly scaled by $U(0.1, 1.0)$.

Results. Table 5 summarizes the pairwise similarities, and Figure 10 visualizes their distributions.

Table 5: Validation of Theorem 1 on CIFAR-10. "Intra" indicates pairs within the same predictive-bias regime, while "Cross" denotes pairs with different biases.

Regime	Comparison	Row-wise Sim	Flat Sim	Prediction (Theorem 1)
Intra-cluster (Same Bias)	G2 Intra (Low-α, dom=3)	$\mathbf{0.989 \pm 0.009}$	0.925 ± 0.083	$\to 1$ (Strong bias, $\epsilon_{LQ} \to 0$) ✓
	G3 Intra (Sub-cluster, dom=6)	$\mathbf{0.990 \pm 0.007}$	0.917 ± 0.101	
	G3 Intra (Sub-cluster, dom=9)	$\mathbf{0.996 \pm 0.002}$	0.963 ± 0.036	
	G1 Intra (High-α, balanced)	0.493 ± 0.138	0.246 ± 0.229	> 0 (Weak bias, larger ϵ_{LQ}) ✓
Inter-cluster (Diff Bias)	G3 Cross (dom=3 $\leftrightarrow$ dom=6)	$\mathbf{-0.228 \pm 0.022}$	-0.223 ± 0.000	< 0 (Misaligned dominant) ✓
	G3 Cross (dom=3 $\leftrightarrow$ dom=9)	$\mathbf{-0.203 \pm 0.037}$	-0.414 ± 0.000	
Quantity Shift (Invariance)	G4 Intra (Quantity-shifted)	0.572 ± 0.150	0.291 ± 0.222	$\approx$ G1 Intra (Scale absorbed) ✓
	G1 $\leftrightarrow$ G4 (Cross-scale)	0.539 ± 0.157	0.294 ± 0.256	

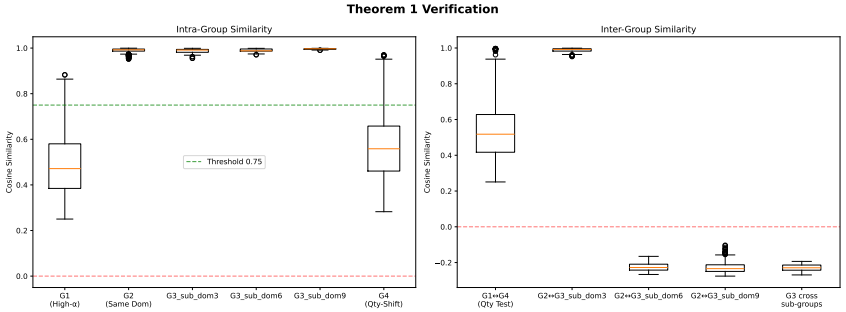

Figure 10: Distribution of row-wise CAM similarities on CIFAR-10. **Left:** Strong predictive bias (G2, G3-subs) forces near-perfect intra-cluster alignment, while weak bias (G1) exhibits wider variance. **Right:** Cross-dominant pairs (G3-cross) strictly separate into negative similarities, whereas quantity shift (G1$\leftrightarrow$G4) remains invariant, mirroring the G1 baseline.

Analysis. The empirical observations perfectly map to the theoretical claims of Theorem 1:

1. **Dominant class dictates alignment.** Clients sharing the same dominant class (G2, G3 sub-clusters) achieve near-perfect similarity ($\geq$0.98). The dominant residual $\|\mathbf{r}_c\|$ suppresses empirical noise, validating the bound $\langle \mathbf{v}_c^{(i)}, \mathbf{v}_c^{(j)} \rangle \geq 1 - \epsilon_{LQ}$ with $\epsilon_{LQ} \to 0$.
2. **Orthogonality across different biases.** Despite sharing the exact identical label set $\{3, 6, 9\}$, cross-dominant pairs in G3 yield strict negative similarities (~ -0.22). This confirms that the predictive bias (local marginal), rather than the label set itself, governs the CAM geometry.
3. **Invariance to quantity shift.** The G1$\leftrightarrow$G4 inter-group similarity (0.539) seamlessly matches the G1 intra-group baseline (0.493). Row normalization successfully absorbs the global gradient scaling induced by heterogeneous dataset sizes.
4. **Noise sensitivity under weak bias.** Balanced clients (G1) exhibit moderate similarity ($\sim$0.49). Without a dominant class to anchor the residual vector, the empirical noise $\xi_T(\delta)$ becomes non-negligible, naturally widening the theoretical error margin ϵ_{LQ}.

B.1.2 Validation of Theorem 2: Feature Skew Robustness

Setup. To validate topological robustness against feature skew, we deploy the PACS domain generalization benchmark (C=7) and partition 50 clients into five 10-client groups:

- **G1–G4 (Same Bias, Varying Domains):** All share Label Pattern A (classes 0,1,2 dominant) but span varying feature domains: Photo (G1, baseline reference), Art (G2, light skew), Cartoon (G3, moderate skew), and Sketch (G4, severe structural skew).
- **G5 (Control: Same Domain, Diff Bias):** Uses the exact same Photo domain as G1, but with a different Label Pattern B (classes 4,5,6 dominant) to isolate the effect of predictive bias.

Results. Table 6 reports the cross-domain similarities with respect to the G1 baseline. Figure 11 illustrates the heatmap and the degradation curve across domain shift severities.

Table 6: Validation of Theorem 2 on PACS. Cross-domain row-wise cosine similarities evaluated against the baseline reference group (G1: Photo, Pattern A).

Regime	Comparison vs. G1	Row-wise Sim	Flat Sim	Prediction (Theorem 2)
	G1 (Photo, Pat A) Intra-group	0.827	0.839	Baseline (Reference)
Intra-bias	G2 (Art, Pat A)	0.706	0.653	
(Feature Skew)	G3 (Cartoon, Pat A)	0.753	0.702	$\geq 1-\epsilon_F$ (Topology preserved) ✓
	G4 (Sketch, Pat A)	0.714	0.637	
Inter-bias (Label Skew)	G5 (Photo, Pat B)	**0.181**	0.134	$\ll 1-\epsilon_F$ (Orthogonal shift) ✓

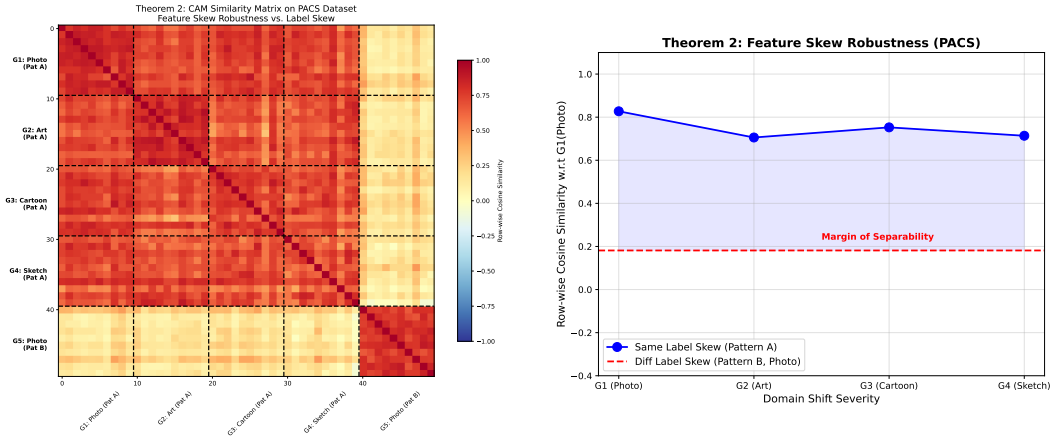

Figure 11: Empirical validation of feature skew robustness. **Left:** CAM similarity heatmap. G1–G4 form a high-similarity block despite spanning four distinct domains, whereas the label-skew control (G5) is strictly isolated. **Right:** Degradation curve. Feature skew induces only a mild similarity drop, maintaining a massive margin of separability above the differing-bias threshold (red dashed line).

Analysis. The empirical results perfectly align with the theoretical bounds of Theorem 2:

1. **Topological robustness to feature skew.** Even under severe structural corruption where texture and color are lost (G4 Sketch), similarity remains remarkably high (0.714). As long as the semantic kernel margin holds ($\kappa_{\text{in}} - \kappa_{\text{out}} > 2\varepsilon_f$), feature skew merely dampens attribution magnitudes without violating the non-negative orthant prior, preserving the CAM topology.
2. **Predictive bias dominates feature space.** Despite sharing the identical feature distribution (Photo) with G1, G5's differing label skew causes similarity to plummet to 0.181. This decisively confirms that CAM geometry is primarily dictated by the predictive bias.
3. **Safe margin for clustering.** As shown in Figure 11 (Right), a massive numerical gap ($\Delta \approx 0.53$) separates the worst-case feature skew (Sketch: 0.714) from the label skew control (G5: 0.181). This robust margin guarantees that cosine-based algorithms can reliably cluster clients by label distributions, immune to local domain variations.

B.1.3 Validation of Theorem 3: Concept Shift Separability via Magnitude Amplification

Setup. To validate the magnitude-amplification mechanism (Theorem 3), we construct a strict concept-shift benchmark on CIFAR-10. We partition 50 clients into five 10-client groups. To strictly isolate concept shift, G1, G4, and G5 share the *exact same* feature distributions $P(x)$ and apparent label marginals $P(y)$, differing only in the latent conditional $P(y \mid x)$:

- **G1 (Normal Reference):** Identity mapping $y = x$ on classes $\{0..4\}$.
- **G4 & G5 (Concept Shift):** Cyclic permutations $y = (x + 1) \bmod 10$ and $y = (x + 2) \bmod 10$.
- **G2 (Feature Control) & G3 (Label Control):** G2 uses disjoint feature support $\{5..9\}$ with $y = x$; G3 enforces label skew (class 2 dominant) to provide baseline comparisons for other shifts.

Results. Table 7 reports the CAM divergences. Under uniform predictive bias, concept shift is detected through a clear *magnitude amplification* in row-normalized Frobenius distance, confirming the bound in (37).

Table 7: Validation of Theorem 3 on CIFAR-10. Comparisons against the normal reference (G1). Frobenius distance serves as the primary metric to capture magnitude amplification.

Regime	Comparison vs. G1	Row-Cos	Flat-Cos	Frob-Dist	Prediction (Theorem 3)
Baseline	G1 (Normal A) Intra-group	0.906	0.902	0.439	Reference (Noise floor)
Concept Shift (Strict)	**G4 (CS +1)**	0.688	0.698	**0.772**	$\geq A_c - B_c$ (Magnitude amplified) ✓
	G5 (CS +2)	0.626	0.622	**0.867**	
Auxiliary Controls	G2 (Feature Control)	0.580	0.311	1.173	Separated (Theorem 2) ✓
	G3 (Label Control)	0.269	0.136	1.314	Separated (Theorem 1) ✓

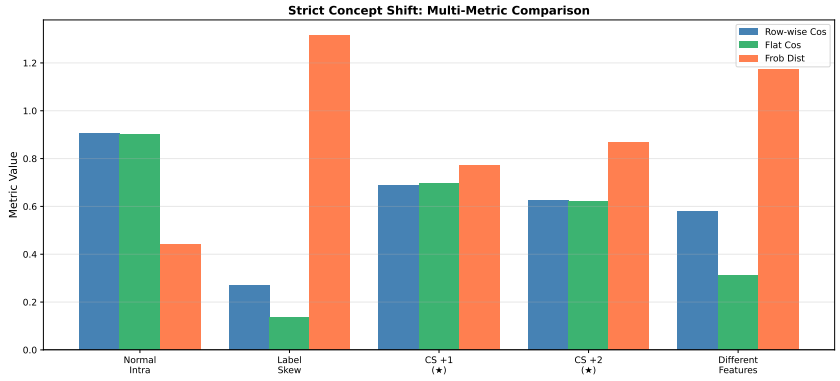

Figure 12: Empirical validation of strict concept shift. The Frobenius distance reveals a clear structural hierarchy: concept-shifted groups (G4, G5) exhibit significant magnitude amplification compared to the G1 reference, scaling monotonically with the permutation severity ($+2 > +1$). Label and feature controls (G2, G3) show even larger, distinct divergences.

Analysis. The empirical results provide tight numerical confirmation of the mechanisms in Theorem 3:

1. **Necessity of magnitude amplification.** Because G1, G4, and G5 possess uniform predictive bias ($\mathbf{p} \approx \frac{1}{C}\mathbf{1}$), the residual vectors lack the dominant direction required to force a cosine sign flip. Instead, concept shift manifests as kernel-amplified *magnitude*. The Frobenius distance jumps from 0.439 (G1-intra) to 0.772 (G4) and 0.867 (G5). The $+2$ cyclic shift accumulates more semantic mismatches per row, correctly producing a larger $\kappa_{\text{in}} - \kappa_{\text{out}}$ amplification than $+1$.
2. **Hierarchical separability across heterogeneities.** Theorem 3 predicts a distinct structural signature for concept shift. The empirical Frobenius distances form a clean hierarchy: **Normal** (0.44) < **Concept Shift** $(0.77$–$0.87)$ < **Feature Skew** (1.17) < **Label Skew** (1.31). This confirms CAM's capacity as a unified diagnostic tool capable of disentangling all distribution shifts.
3. **Robust empirical margin for clustering.** The observed separation margin for concept shift ($\Delta \approx 0.33$–0.43) comfortably exceeds the intra-group noise fluctuation estimated from G1 ($2\xi_T(\delta) \leq 0.05$). This satisfies the empirical margin condition (36), ensuring that concept-shifted clients reliably form internal cohesive sub-clusters (e.g., G4 vs. G5 Frob $=0.779$) without relying on restrictive sign-reversal hypotheses.

B.2 Experimental Details

Implementation Details. We implement the experiments in a PyTorch-based simulation environment, building upon the open-source pFL benchmark [58]. We employ the mini-batch SGD optimizer with a learning rate of 0.01. Unless explicitly specified, the local batch size is set to 32, and the number of local training epochs is 10. We simulate a total of 100 clients with a 20% participation ratio per round, training for 200 communication rounds. For FedCap, we execute 1 round during the warm-up phase to acquire the CAMs. We run all experiments on Intel Xeon Gold 6133 CPU, RTX4090 GPU.

Hyperparameters and Threshold Search. For negligible row norms, the normalized row is zero; the cosine argument is clipped to $[-1, 1]$. An all-zero client signature has zero cosine with other clients, and self-distances are set to zero. We rank the $N(N-1)/2$ off-diagonal angular distances, use average ranks for ties, and map rank r to $(r-1)/(N(N-1)/2-1)$. Optional sparse transmission sends nonzero rows with class indices for reconstruction in fixed order; the dense payload calculation below does not assume this encoding.

The defaults are specified in the main experimental setup. The implementation uses average-linkage hierarchical clustering on the rank-normalized dissimilarity matrix. It evaluates 1000 equally spaced thresholds in $[0.1, 1.0]$, retains partitions with $3 \leq K \leq 20$, and minimizes the conflict-aware objective. If no candidate is admissible, it scans 300 thresholds in $[0.01, 3.0]$ and selects the partition whose cluster count is closest to 11. This fallback is a cluster-count heuristic rather than minimization of the conflict-aware objective. The row-normalization tolerance is 10^{-8}, and the supply-demand construction uses a zero demand threshold.

Communication and Client-Side Cost. With float32 CAM entries and class proportions, and a bit-packed class-presence mask, the one-time clustering metadata payload per client is $4C^2+4C+\lceil C/8 \rceil$ bytes. This is a theoretical payload count under the stated encoding, not a measured network cost; it excludes protocol headers and any additional data-size message for aggregation. Public-reference distribution and ordinary model transfers are separate costs. After warm-up, the CAM and class metadata are not recomputed or resent. Each training round uses the client's own cluster model and, when available, a server-averaged supplier target. Sending one averaged target need not increase the client's payload linearly with the number of suppliers. CAM extraction adds validation-sample forward/backward processing and ghost inner products during warm-up. Avoiding explicit per-sample parameter gradients does not make this computation free. The server timing table does not measure these client-side or end-to-end costs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sections 1 and 4 explicitly present the main contributions of this work. The abstract and introduction accurately reflect both the proposed theoretical framework and the scope of empirical validation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the limitations of this work. Furthermore, in Section 4, we investigate the computational cost of the clustering method under large-scale client settings, demonstrating its favorable scalability.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results presented in this paper are accompanied by complete assumptions and rigorous proofs in the appendix. In particular, the theoretical foundations of the class-interaction attribution matrix for clustering under three types of distribution shifts are provided. Detailed assumptions and proofs for Theorems 1–3 can be found in Appendix A.4, while the convergence analysis of FedCap is presented in Appendix A.6.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details and hyperparameter settings are comprehensively described in Section 4 and Appendix B.2. These details ensure that the main experimental results can be reproduced even without direct access to the original code and datasets.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and datasets used in this work are publicly available via the anonymous link `https://anonymous.4open.science/r/FedCap-605E`, along with additional resources in the supplementary material. These materials are sufficient to support the reproduction of the main results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and testing details are clearly specified in Section 4 and Appendix B.2, including data splits, hyperparameters, and optimization settings. While some hyperparameter choices are heuristic, we additionally provide empirical analyses demonstrating the impact of different hyperparameter configurations.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The reported error bars are obtained from three independent experimental runs. The variability reflected by the error bars arises from stochastic factors during training, such as SGD randomness and client sampling at each round, rather than changes in the dataset. Therefore, the reported results provide meaningful statistical significance.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 specifies that we run all experiments on Intel Xeon Gold 6133 CPU, RTX4090 GPU.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All research complies with NeurIPS Ethics Guidelines. No human subjects or sensitive personal data involved. Benchmark datasets used contain publicly available non-identifiable information.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The proposed method is primarily focused on improving model training in federated learning systems. At this stage, we do not identify any direct societal impacts, either positive or negative. The work is methodological in nature and does not involve applications that could raise immediate ethical, social, or policy concerns.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not introduce or release any new datasets or pre-trained models with a high risk of misuse. The contributions are limited to methodological developments and experimental evaluations based on publicly available datasets. Therefore, no specific safeguards are required for responsible release.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and assets used in this work are publicly available and properly credited. For example, the CIFAR datasets are released under the CC-BY 4.0 license, and other datasets such as PACS are used in accordance with their respective terms of use. Relevant citations and license acknowledgments are provided in the references to ensure compliance with usage requirements.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or models released beyond methodological implementations.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subject research conducted. All data sources consist of publicly available benchmark corpora.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subject research conducted. All data sources consist of publicly available benchmark corpora.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs not used in methodology development or experimental analysis. Grammarly employed solely for grammar checking in manuscript preparation.