Disentangling Federated Learning Heterogeneity: A Dual-Perspective Analysis of Quantifying Skew versus Scarcity

Wenkai Zeng[1] **Nan Yang**[1*] **Zhiyu Zhu**[2] **Zhibo Jin**[2] **Dong Yuan**[1*]
The University of Sydney[1] University of Technology Sydney[2]

Abstract

Federated Learning faces significant challenges due to data heterogeneity, which manifests as Label Distribution Skew and label missingness. We propose Skew-Scarcity Disentanglement Indicator (SSDI), a novel metric that decomposes heterogeneity into two disentangled components: Label Distribution Skew (LDS) (quantity skew of present labels) and Label Coverage Deficiency (LCD) (deviation due to missing labels). Using a PAC-Bayesian framework, we derive a generalization bound indicating that Label Coverage Deficiency becomes the dominant risk factor as the number of clients increases, severely degrading accuracy on rare labels. Our study reveals that, for a fixed number of labels, increasing clients is a primary driver of per-label accuracy variance by exacerbating Label Coverage Deficiency. Moreover, a higher global missing rate intensifies this divergence effect and can precipitate severe performance breakdown at a lower critical threshold of clients. Experiments on vision benchmarks confirm that SSDI accurately captures the severity of performance divergence. The SSDI framework provides a principled tool for diagnosing heterogeneity and guiding targeted mitigation strategies. The code for the SSDI-controlled client-label matrix generation used in our experiments is available at https://github.com/wkzeng/SSDI.git.

* Corresponding authors: Nan Yang and Dong Yuan. (n.yang@sydney.edu.au and dong.yuan@sydney.edu.au)

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Federated Learning has emerged as a foundational paradigm for training machine learning models across decentralized data sources while preserving data privacy (McMahan et al., 2017). However, its practical deployment faces fundamental challenges due to data heterogeneity—the statistical divergence in data distributions across clients (Li et al., 2020; Konečný et al., 2016). While this challenge manifests empirically through performance degradation and convergence issues (Li et al., 2020), existing theoretical frameworks lack the granularity to precisely diagnose and quantify the distinct mechanisms through which heterogeneity impacts learning.

Theoretical Gap in Heterogeneity Modelling. Current theoretical analyses of federated learning predominantly treat heterogeneity as a monolithic challenge, failing to distinguish between two fundamentally different phenomena: *Label Distribution Skew* (variations in class proportions across clients) and *label missingness* (complete absence of categories from client datasets) (Zhu et al., 2021; Lu et al., 2024). This conflation obscures their distinct effects on model performance—while distribution skew leads to imbalanced learning, missingness can cause complete failure on rare categories (Shuai et al., 2022). Moreover, existing PAC-Bayesian frameworks employ static priors that cannot adapt to varying heterogeneity patterns (McAllester, 1999), missing the opportunity to precisely quantify how different disparity types contribute to generalization risk.

Limitations of Current Heterogeneity Metrics. Existing approaches to measure non-IIDness suffer from critical limitations (Domini et al., 2026). Some methods focus solely on distribution divergence metrics without considering label missingness (Hsu et al., 2019). Others quantify heterogeneity through performance-based measures that conflate multiple effects (Yang et al., 2021). Most critically, no existing metric provides the orthogonal decomposition needed to guide targeted mitigation strategies (Shi et al., 2024).

Our Approach and Contributions. We address these limitations through a dual-disparity perspective that mathematically disentangles heterogeneity into orthogonal components. Our work makes four key contributions:

(1) **SSDI: Skew-Scarcity Disentanglement Indicator:** A theoretically grounded metric that decomposes federated learning heterogeneity into Label Distribution Skew (quantifying the non-uniform distribution of present labels) and the Label Coverage Deficiency (quantifying the deviation due to missing labels).

(2) **SSDI-Modulated PAC-Bayesian Framework:** We design an adaptive prior whose strength is dynamically regulated by SSDI, allowing the derivation of a generalization bound that explicitly quantifies how each disparity component influences risk.

(3) **Critical Scaling Analysis:** We identify a critical condition $K \sim \kappa \cdot C \ln C$ where dominance shifts from LDS to LCD disparity, explaining why performance divergence intensifies with client fragmentation.

(4) **Empirical Validation:** Experiments systematically verify our theoretical predictions, showing SSDI's superiority in capturing performance divergence patterns.

Our framework provides the first unified approach that not only measures heterogeneity but also characterizes its performance implications through rigorous theoretical analysis, offering researchers a principled tool for diagnosing federated learning challenges.

2 RELATED WORK

Our work intersects with four main research threads: federated optimization under heterogeneity, theoretical analyses of federated learning, heterogeneity quantification methods, and personalized federated learning.

Federated Optimization under Heterogeneity. The field was established by FedAvg (McMahan et al., 2017), which suffers from client drift under non-IID data. Subsequent algorithms address this through various strategies: FedProx (Li et al., 2020) introduces a proximal term to constrain local updates; SCAFFOLD (Karimireddy et al., 2020) employs control variates to correct client drift; and adaptive optimization methods (Reddi et al., 2021) extend adaptive optimizers to federated learning settings. While these methods provide algorithmic solutions, they lack fine-grained diagnostic capabilities to determine which type of heterogeneity is most impactful in a given scenario.

Theoretical Analyses of Federated Learning. Significant effort has been devoted to understanding federated learning convergence and generalization. Early work provided optimization guarantees (Konečný et al., 2016), with subsequent studies analyzing FedAvg convergence on non-IID data (Li et al., 2020) and generalization gaps between centralized and decentralized federated learning (Sun et al., 2026, 2023). The PAC-Bayesian framework has been applied to provide generalization bounds (McAllester, 1999; He et al., 2019) and analyze algorithmic stability (Hardt et al., 2016). However, these analyses typically treat heterogeneity as a uniform noise term, failing to distinguish between distribution skew and missingness effects.

Heterogeneity Quantification Methods. Existing approaches to measure non-IIDness suffer from critical limitations. Some methods focus solely on distribution divergence metrics without considering label missingness (Zhu et al., 2021). Others quantify heterogeneity through performance-based measures that conflate multiple effects (Yang et al., 2021). Most critically, no existing metric provides the orthogonal decomposition that our SSDI offers, as summarized in Table 1. Recent benchmarks like ProFed (Domini et al., 2026) highlight the need for more nuanced heterogeneity evaluation frameworks.

Personalized Federated Learning. Recognizing the limitations of single global models, personalized federated learning approaches have gained traction. Ditto (Li et al., 2021) learns personalized models with fairness constraints, while pFedMe (T Dinh et al., 2020) uses Moreau envelopes for personalization. More recent methods like pFedAFM (Yi et al., 2024) address batch-level heterogeneity through adaptive feature mixture. These approaches implicitly handle heterogeneity but lack explicit quantification of its components.

Emerging Directions. Several emerging directions are particularly relevant to our work. Self-supervised learning in federated learning settings (Yang et al., 2023; Wang et al., 2023) addresses label scarcity but does not explicitly distinguish heterogeneity types. Asynchronous federated learning methods (Iakovidou & Kim, 2024; Kang & Li, 2024) handle system heterogeneity but often overlook statistical heterogeneity decomposition. Bayesian approaches like FedBE (Chen and Chao, 2021) provide robust aggregation but lack fine-grained heterogeneity analysis. While these approaches address various aspects of federated learning heterogeneity, they fail to provide the fine-grained decomposition needed for precise diagnosis and targeted mitigation.

Positioning Our Work. Unlike prior approaches that either treat heterogeneity monolithically or provide only coarse-grained measurements, our SSDI frame-

work offers a principled decomposition that directly informs both theoretical analysis and practical mitigation strategies. By embedding this decomposition within a PAC-Bayesian framework, we provide the first unified approach that connects heterogeneity measurement to generalization guarantees. **Our work differs from existing literature in three key aspects:**

(a) We provide the first fine-grained decomposition of heterogeneity into orthogonal components with distinct performance implications.

(b) We derive a theoretically grounded metric that directly informs generalization bounds.

(c) We identify critical scaling conditions that predict when different heterogeneity types dominate performance degradation.

Table 1: Comparison of heterogeneity quantification methods. DSM (Distinguishes Skew/Missingness), PPD (Predictive of Performance Divergence)

Metric	DSM	PPD
Earth Mover's Distance	×	×
Dirichlet-α	×	×
Performance Variance	×	✓
SSDI (Ours)	✓	✓

We emphasize that SSDI is not designed to replace prior metrics, but rather to complement them by providing a mechanism-level diagnostic tool. The two SSDI components exhibit distinct scaling behaviors and system-level implications that cannot be recovered from a single aggregated heterogeneity measure.

3 THE SKEW-SCARCITY DISENTANGLEMENT INDICATOR

3.1 Problem Formulation and Notation

Consider a federated learning system with K clients and C distinct classes. Let $n_{k,c}$ denote the number of samples of class c on client k. The local data volume at client k is defined as $n_k = \sum_{c=1}^{C} n_{k,c}$, while the global class size is $N_c = \sum_{k=1}^{K} n_{k,c}$. The overall dataset size is $N_{\text{total}} = \sum_{k=1}^{K} n_k = \sum_{c=1}^{C} N_c$. The local class proportion is $\rho_{k,c} = n_{k,c}/n_k$, and the global class proportion is $\rho_c^{\text{global}} = N_c/N_{\text{total}}$.

3.2 Dual-Viewpoint Deviation Framework

The core idea of SSDI is to characterize data heterogeneity from two complementary viewpoints, capturing distinct but interacting sources of non-IID behavior in federated learning.

3.2.1 Label Coverage Deficiency (LCD)

Label Coverage Deficiency (LCD) captures the *structural pattern* of label distribution across clients by measuring how unevenly samples of each class are distributed throughout the system. Unlike local distribution skew, LCD explicitly accounts for both partial scarcity and complete absence of labels on certain clients.

$$d^c = \sqrt{\frac{1}{N_c} \cdot \frac{1}{N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}. \quad (1)$$

The quantity d^c measures the unevenness of class c across clients. High values arise when a class is concentrated on a small subset of clients or completely missing from many others.

Relation to quantity shift. LCD is related to classical quantity shift, but goes beyond aggregate sample imbalance by explicitly characterizing the *structural coverage pattern* of labels at the class level. In particular, LCD captures scenarios where labels are systematically absent from a large fraction of clients, which cannot be inferred from global class counts alone.

Systemic implication. A high LCD reflects a *label missingness accumulation effect*, where the absence of certain labels from many clients prevents the global model from receiving sufficient gradient signals. As the number of clients increases, this effect can dominate learning dynamics and lead to catastrophic performance collapse on under-covered classes.

3.2.2 Label Distribution Skew (LDS)

Label Distribution Skew (LDS) quantifies the deviation of local label distributions within each client from the global label distribution, while accounting for the actual presence of labels.

$$d_k = \sqrt{\frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}. \quad (2)$$

The quantity d_k measures how much the label distribution of client k deviates from the global distribution.

Relation to label shift. As defined above, LDS is closely related to the classical *label shift* assumption, where client-level label distributions $P_k(y)$ differ across clients while the class-conditional feature distributions $P(x \mid y)$ remain consistent.

Systemic implication. A high LDS signifies *systemic long-tail bias*, where non-uniform label distributions induce biased yet non-zero gradients that favor frequently observed labels over rare ones. This results in client drift and biased aggregation, even when all labels are globally present.

3.3 Constructing the SSDI Metrics

The class-level LCD deviations and client-level LDS deviations are concatenated into a unified deviation vector:

$$\mathbf{v} = \left(d^1, d^2, \ldots, d^C, d_1, d_2, \ldots, d_K\right)^\top. \tag{3}$$

The Skew-Scarcity Disentanglement Indicator (SSDI) is defined as:

$$\text{SSDI} = \frac{\|\mathbf{v}\|_2}{\|\mathbf{v}_{\max}\|_2}, \tag{4}$$

where $\|\mathbf{v}_{\max}\|_2$ denotes a (C, K)-adaptive theoretical normalization constant used to bound the range of SSDI $\in [0, 1]$.

3.4 Matrix Formulation and Interpretation

The SSDI admits an equivalent matrix formulation that provides a joint distribution interpretation; full derivations are given in Appendix B.2.

Empirical joint distribution. Let $P \in \mathbb{R}^{C\times K}$ with entries $p_{c,k} = n_{k,c}/N_{\text{total}}$.

Ideal independent distribution. Let $Q \in \mathbb{R}^{C\times K}$ with entries $q_{c,k} = \rho_c^{\text{global}} \cdot (n_k/N_{\text{total}})$.

Deviation matrix. The deviation matrix $D = P - Q$ measures the departure from the ideal IID case.

Weighted norm formulation. The SSDI deviation magnitude can be expressed as:

$$\|\mathbf{v}\|_2 = \|W \odot D\|_F, \tag{5}$$

where $\odot$ denotes the Hadamard product and $\|\cdot\|_F$ is the Frobenius norm. The weight matrix W is defined as:

$$w_{c,k} = \sqrt{\frac{N_{\text{total}}}{N_c} + \frac{N_{\text{total}}}{n_k}}. \tag{6}$$

Interpretation of the weighting scheme. The term $\frac{N_{\text{total}}}{N_c}$ assigns larger weights to rare classes, reflecting the fact that uneven distribution of a rare class poses a greater risk to system-wide generalization than a comparable imbalance of a frequent class. This design ensures SSDI remains sensitive to long-tail effects. The term $\frac{N_{\text{total}}}{n_k}$ assigns larger weights to small-capacity clients, whose local data distributions are more prone to overfitting and client drift. Together, these weights balance the contributions of label scarcity and client imbalance in heterogeneity assessment.

3.5 Normalization and Extreme Case Analysis

We adopt a (C, K)-adaptive normalization based on the theoretical supremum over the unconstrained joint-distribution space, where some classes or clients may approach zero mass in the limiting extremal configuration.

Let $m = \min(C, K)$. The limiting extremal structure is permutation-type when $C = K$, and otherwise follows a rectangular matching skeleton of size m. The resulting extremal deviation matrix is

$$(D_{\max})_{c,k} = \begin{cases} \frac{m-1}{m^2}, & c = k \le m, \\ -\frac{1}{m^2}, & c \ne k,\ c \le m,\ k \le m, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

with active-block weight

$$(W_{\max})_{c,k} = \sqrt{2m}, \qquad c \le m,\ k \le m. \tag{8}$$

Hence

$$\|\mathbf{v}_{\max}\|_2 = \|W_{\max} \odot D_{\max}\|_F = \sqrt{2\left(1 - \frac{1}{\min(C, K)}\right)}. \tag{9}$$

Detailed derivations are given in Appendix B.2.

Properties. SSDI $= 0$ corresponds to IID data, and SSDI $= 1$ to the limiting theoretical extremal configuration under (C, K). Larger SSDI indicates stronger overall heterogeneity.

4 THEORETICAL ANALYSIS

4.1 PAC-Bayesian Framework

To provide a rigorous theoretical foundation for the SSDI, we embed it within a PAC-Bayesian framework. This allows us to derive a generalization bound that explicitly quantifies how the two components of SSDI influence the global model's risk.

Prior Distribution Design. We design an adaptive prior distribution $P(\boldsymbol{\theta}) = \mathcal{N}(0, \boldsymbol{\Sigma}_p)$ whose strength is dynamically regulated by the SSDI metric. Specifically, the inverse covariance matrix is given by $\boldsymbol{\Sigma}_p^{-1} = \gamma \cdot \text{Diag}(\boldsymbol{\omega})$, $\boldsymbol{\omega} = \phi(\text{SSDI}) \cdot \boldsymbol{\omega}_0$, where parameter $\gamma > 0$ serves as a global regularization strength coefficient that controls the overall influence of the adaptive prior,

$$\phi(\text{SSDI}) = \frac{\text{SSDI}}{1 - \text{SSDI} + \epsilon} \tag{10}$$

is a heterogeneity strength function that amplifies the prior's influence as SSDI increases,

$$\boldsymbol{\omega}_0 = \left(\frac{N_{\text{total}}}{N_1}, \ldots, \frac{N_{\text{total}}}{N_C}, \frac{N_{\text{total}}}{n_1}, \ldots, \frac{N_{\text{total}}}{n_K}\right)^T \tag{11}$$

is a baseline weight vector that applies stronger regularization to rare classes (small N_c) and small clients (small n_k).

Posterior Distribution Modelling. The client-local and global learning processes induce a posterior distribution $Q(\boldsymbol{\theta}) = \mathcal{N}(\mu, \boldsymbol{\Sigma}_q)$ (where $\boldsymbol{\Sigma}_q = \text{Diag}(\sigma_{q,1}^2, \ldots, \sigma_{q,d}^2)$) characterized by parameter drift and model uncertainty, with parameter constraints (KL-divergence boundary):

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\|^2 \leq \frac{2}{\lambda_{strong}} D(Q\|P) \quad , \tag{12}$$

where $\boldsymbol{\theta}^{(t)}$ denotes the global model parameters at the beginning of the t^{th} training round.

Definition 1 (Posterior Parameter Interpretation)**.** The posterior parameters admit the following interpretation: the posterior mean μ captures parameter drift relative to the prior, while the posterior variance $\sigma_{q,i}^2$ captures uncertainty along parameter dimension i. Larger mean drift $\|\mu\|_2^2$ indicates stronger systematic deviation, whereas larger posterior variances imply higher generalization risk.

Assumptions. To gain deeper insight into how system parameters (number of clients K, number of classes C, overall dataset size N_{total}) affect heterogeneity risk, we define the label missingness indicator $M_{c,k} = \mathbb{I}(n_{k,c} = 0)$, which equals 1 if client k has no samples of class c, and 0 otherwise. We then conduct asymptotic analysis under the following statistical assumptions; their motivation is discussed in Appendix B.8.

Assumption 1. *Client data sizes follow a Pareto distribution: $n_k \sim Pareto(x_m, \alpha_{pareto})$, where the shape parameter $\alpha_{pareto} = 2$ and the scale parameter $x_m = N_{total}/(2K)$.*

Assumption 2. *Global label distribution follows Zipf's law: $N_c \sim Zipf(\beta_{zipf})$, where the exponent parameter $\beta_{zipf} = 1$.*

Assumption 3. *Effective parameter drift at the client level is proportional to label missingness: $\mu_k^2 \propto \sum_c M_{c,k} N_c$. This implies that missing a label with larger global prevalence (larger N_c) induces more severe parameter drift on affected clients.*

Assumption 4. *Posterior variance is inversely proportional to local data size: $\sigma_{q,k}^2 \propto n_k^{-1}$. This captures the intuition that clients with smaller local datasets exhibit higher model uncertainty.*

4.2 Generalization Error Bound

Under the above setup, we derive the following main theorem for the global model's generalization error $R(Q)$.

Theorem 1 (SSDI-dependent Generalization Bound)**.** *For any data distribution and any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the generalization error of the global model satisfies:*

$$R(Q) \leq \hat{R}(Q) + \sqrt{\frac{\Gamma_{LDS} + \Gamma_{LCD} + \Psi + \log \frac{1}{\delta}}{2N_{total} - 1}}, \tag{13}$$

*where $\hat{R}(Q)$ is the empirical risk. Term Ψ is a lower-order model complexity term that captures the intrinsic complexity of the hypothesis space through the log-determinant of the prior precision matrix and the entropy of posterior uncertainties. The key heterogeneity-dependent penalty terms are Γ_{LDS} **(label distribution disparity penalty)** and Γ_{LCD} **(label coverage disparity penalty)**. This naming follows the role of class-level rarity in skew-induced imbalance and client-level data sparsity in coverage deficiency.*

$$\Gamma_{LDS} = \frac{\gamma}{2} \phi(SSDI) \left[\sum_{c=1}^{C} \frac{N_{total}}{N_c} (\sigma_{q,c}^2 + \mu_c^2) \right], \tag{14}$$

$$\Gamma_{LCD} = \frac{\gamma}{2} \phi(SSDI) \left[\sum_{k=1}^{K} \frac{N_{total}}{n_k} (\sigma_{q,k}^2 + \mu_k^2) \right]. \tag{15}$$

Here, μ_c^2, μ_k^2 and $\sigma_{q,c}^2, \sigma_{q,k}^2$ denote effective class-level and client-level drift and uncertainty terms in the heterogeneity penalty decomposition.

Corollary 1 (Amplification Effect of SSDI)**.** *The function $\phi(SSDI)$ multiplicatively amplifies the entire risk penalty. As the overall heterogeneity (SSDI) increases, the generalization bound yields a larger upper bound, indicating higher expected error. This amplification effect follows directly from the construction of the adaptive prior and scales monotonically with the measured heterogeneity.*

Proposition 1 (Decomposition and Impact Mechanism of Heterogeneity Penalty)**.** *The generalization error penalty in Theorem 1 naturally decomposes into two components with distinct physical meanings:*

Γ_{LDS}: Quantifies non-uniform data distribution for present labels, where certain labels have excessive or insufficient samples on certain clients, but all labels have at least some samples. This heterogeneity leads to imbalanced performance across labels, as the model may bias towards labels with more samples.

Γ_{LCD}: Quantifies distribution deviation due to label missingness, where certain labels are completely absent from multiple clients. This heterogeneity causes the model to completely ignore some labels, resulting in very low accuracy for missing labels.

Physical Mechanism Analysis: The two heterogeneity types affect model performance through different mechanisms:

Label Distribution Skew (Γ_{LDS}) is characterized by a gradient of performance across labels: higher accuracy for well-represented labels and lower for under-represented ones—but all labels have some learning opportunity.

Label Coverage Deficiency (Γ_{LCD}), however, triggers a polarized performance pattern: near-zero accuracy for missing labels (model fails to learn their features) alongside relatively normal accuracy for other labels, which significantly increases the variance and range of label-wise accuracy.

Proof Sketch: This decomposition arises from the structure of weight vector $\boldsymbol{\omega}_0$, whose first C components correspond to class weights (N_{total}/N_c) and remaining K components correspond to client weights (N_{total}/n_k). Expanding the KL-divergence according to this structure yields the decomposition.

Detailed derivations are provided in Appendix B.5 and Appendix B.5.3.

4.3 Asymptotic Dominance Analysis

Theorem 2 (Asymptotic Order of Heterogeneity Penalty)**.** *In typical federated learning systems satisfying the above assumptions, the dominant parts of the generalization error penalty satisfy the following asymptotic relations:*

$$\Gamma_{LDS} \sim O\left(C^3 \ln^2 C\right), \quad \Gamma_{LCD} \sim O\left(CK^2\right),$$

where Γ_{LDS} and Γ_{LCD} are the generalization penalty terms derived from the LDS and LCD components, respectively.

Proof Sketch: Substitute the model assumptions into the definitions of Γ_{LDS} and Γ_{LCD}, and compute the expected order of key terms (such as $\sum_k 1/n_k$, $\sum_c 1/N_c$) under Pareto and Zipf distributions.

Corollary 2 (Dominance Critical Condition)**.** *The relative dominance between LDS and LCD is determined by the scaling relationship between the number of clients K and the number of classes C.*

The critical scaling condition occurs when: $K \sim \kappa \cdot C \ln C$, where κ is a constant aggregating system-specific factors.

The asymptotic behaviour exhibits three distinct regimes:

- *When $K \ll \kappa \cdot C \ln C$, LDS dominates the heterogeneity risk, as systemic long-tail bias prevails. In this regime, $\Gamma_{LDS} \sim O(C^3 \ln^2 C)$ dominates over $\Gamma_{LCD} \sim O(CK^2)$.*
- *When $K \gg \kappa \cdot C \ln C$, LCD becomes the dominant risk factor, with missing label effects amplified by client fragmentation. In this regime, $\Gamma_{LCD} \sim O(CK^2)$ dominates over $\Gamma_{LDS} \sim O(C^3 \ln^2 C)$.*
- *Near the critical point $K \approx \kappa \cdot C \ln C$, both heterogeneity types contribute comparably to the overall risk.*

Corollary 3 (Limited Effect of Overall Dataset Size)**.** *When C , K and data distribution are fixed, increasing the overall dataset size N_{total} does not change the dominant asymptotic order of the dominant penalty terms Γ_{LDS} and Γ_{LCD}. This indicates that merely increasing the total data volume cannot mitigate the inherent risk introduced by systemic structural heterogeneity (such as inherent Label Distribution Skew and client data distribution patterns). The main benefit lies in providing more information for the model to learn better representations, which may indirectly reduce μ and σ_q^2, but this effect is indirect and has diminishing returns.*

Detailed asymptotic derivations are provided in Appendix B.6 and Appendix B.7.

4.4 Performance Divergence Mechanism

Corollary 4 (Mechanism of Performance Divergence)**.** *Based on the physical mechanism analysis in Proposition 1, performance divergence is most severe when LCD dominates, because:*

1. ***Extreme accuracy disparity**: Near-zero accuracy for missing labels versus relatively normal accuracy for non-missing labels creates an extreme accuracy range.*
2. ***Variance amplification**: The extreme differences significantly increase cross-label performance variance.*
3. ***Model bias reinforcement**: The model's complete neglect of missing labels is reinforced during aggregation.*

In contrast, when LDS dominates, performance imbalance is relatively moderate as all labels have at least some learning opportunity.

Beyond generalization performance, data heterogeneity also affects the optimization process in federated learning. Our analysis suggests that the client parameter drift $\mathbb{E}\|\Delta_k^{(t)}\|^2$ is related to the heterogeneity-dependent KL penalty induced by the adaptive prior, which directly incorporates SSDI. This provides an optimization

perspective explaining why high-heterogeneity scenarios require more careful client selection and aggregation strategies; further discussion is given in Appendix B.8.

5 EXPERIMENTS

To validate the effectiveness of SSDI and its theoretical implications, we conducted extensive experiments on vision benchmarks under various heterogeneity settings.

5.1 Experimental Setup

Datasets & Models. We employed the MNIST (Lecun et al., 1998), CIFAR-100 (Krizhevsky and Hinton, 2009) and Tiny ImageNet (Deng et al., 2009) datasets, with dataset-specific models including ResNet-18 (He et al., 2016) and ShuffleNetV2 (Ma et al., 2018). Non-IID data were generated under the constraint of a fixed total data volume, by drawing client data sizes from Pareto($\alpha = 2.0$), label distributions from Zipf($\beta = 1.0$), and applying a global label missing rate (MR).

Evaluation Protocol. The primary evaluation metrics included: the evolution of the ratio between the two SSDI components, the **Deficiency-to-Skew Ratio (DSR)** is defined as the ratio between the $\mathbf{SSDI}_{\mathrm{LCD}}$ and $\mathbf{SSDI}_{\mathrm{LDS}}$ components,

$$\mathrm{DSR} = \frac{\mathrm{SSDI}_{\mathrm{LCD}}}{\mathrm{SSDI}_{\mathrm{LDS}}} = \frac{\|W \odot D_{\mathrm{LCD}}\|_F}{\|W \odot D_{\mathrm{LDS}}\|_F}, \tag{16}$$

which reflects the shift in dominance between disparity types; global accuracy, loss, divergence, and the trends and standard deviation of per-class accuracy, which directly measures performance divergence. We also reported the range (max-min) and variance of accuracy across classes.

Heterogeneity Scenarios. We systematically varied the number of clients, the number of classes, and the global missing rate to observe their impact on SSDI and the final model performance. Detailed dataset specifications, training hyperparameters, and the non-IID partitioning procedure are summarized in Appendix C.1, specifically Sections C.1.1, C.1.2, and C.1.3.

5.2 SSDI Behavior and Critical Points Analysis

Figure 1 shows that, for a fixed label count $C = 10$, as the number of clients increases, both the overall SSDI and its individual components exhibit a rapid initial decrease followed by a gradual plateau. This pattern holds consistently across different missing rates, demonstrating SSDI's sensitivity to client fragmentation.

The DSR reveals a more nuanced pattern, exhibiting a U-shaped trend of first decreasing and then increasing, as shown in Figure 2a and Figure 2b. By fitting the DSR curve, we identify the extremum point K_{crit} where the dominance shifts from LDS disparity to LCD disparity. Two key theoretical predictions are empirically validated: (1) K_{crit} decreases with increasing missing rate, and (2) K_{crit} increases proportionally with the number of classes. For instance, when the number of classes changes from 10 to 20, K_{crit} shifts rightward, confirming the proportional relationship $K_{crit} \propto C$. Further analyses of SSDI behavior and critical-point scaling are provided in Appendix C.2 and Appendix C.2.3.

5.3 Performance Divergence Patterns

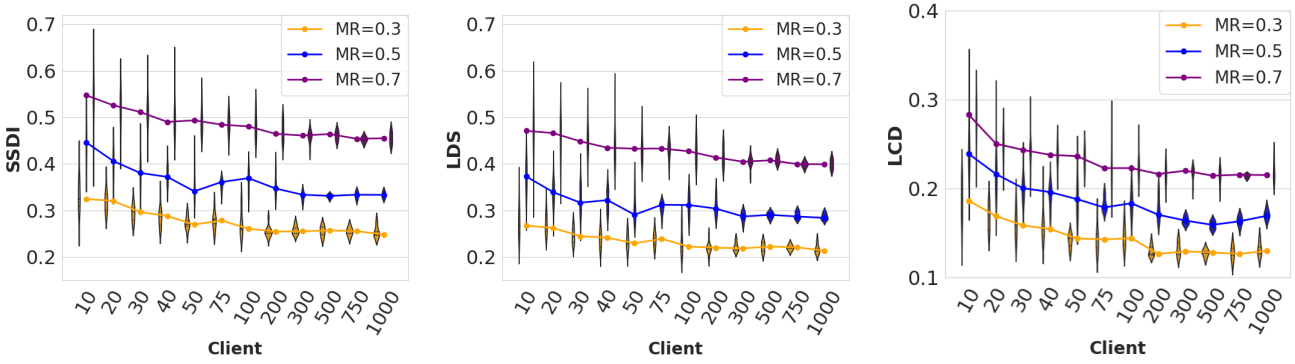
When the number of labels is fixed, the model's performance becomes progressively lower and more uneven as K increases. Figure 3 illustrates this performance divergence phenomenon: as client count grows, rare label accuracy deteriorates significantly while the most frequent labels maintain high accuracy levels.

We systematically recorded accuracy for each class and observed consistent trends: as K increases, global accuracy decreases, class consistency deteriorates, accuracy variance and range increase. This phenomenon reflects the gradual failure in recognizing rare labels after the dominant heterogeneity shifts from LDS disparity to LCD disparity. The effect is further amplified under higher missing rates, with complete analyses of class-wise accuracy, rare-versus-frequent label behavior, and accuracy dispersion provided in Appendix C.3, Sections C.3.1–C.3.3.

5.4 Robustness to Dataset Size Analysis

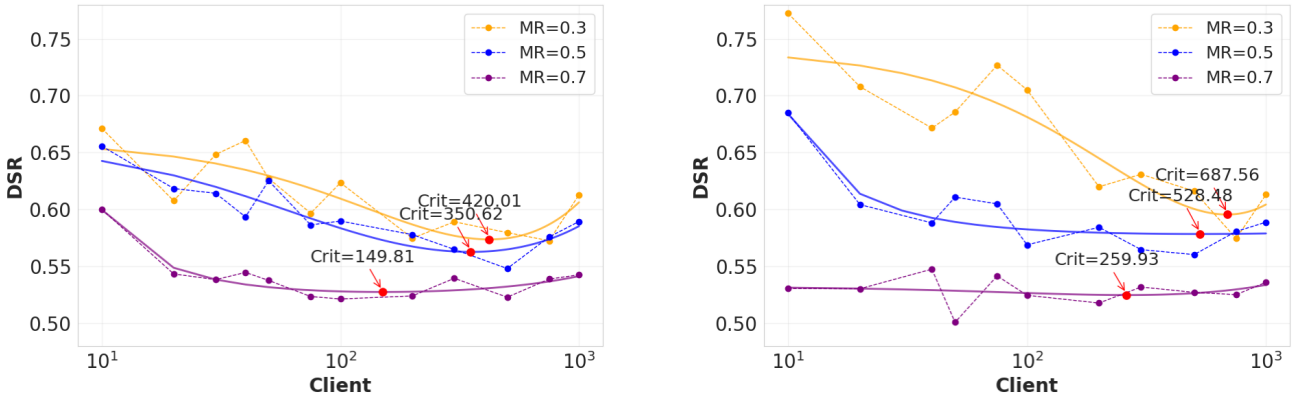
In experiments varying the total sample size, the SSDI-related metrics consistently remained within a reasonable range despite some fluctuations. This demonstrates SSDI's precise characterization of heterogeneity in non-IID data distributions, largely stable with respect to absolute data volume.

Despite the theoretical possibility that a dataset with a larger total sample size could be obtained by proportionally scaling from a smaller distribution, our experiments generated each dataset independently for rigorous validation. As shown in Figure 4, for fixed underlying data distributions satisfying $N \gg C \cdot K$, the influence of total sample size is notably limited, and global accuracy remains relatively stable. Additional performance metrics and robustness evidence are reported in Appendix C.4, Section C.4.1.



(a) SSDI (b) Label Distribution Skew (c) Label Coverage Deficiency

Figure 1: SSDI and its components (LDS, LCD) across different client counts for 10 classes. The rapid initial decrease followed by gradual plateau reflects SSDI's sensitivity to client fragmentation across all missing rates.



(a) Labels = 10 (b) Labels = 20

Figure 2: (a) Deficiency-to-Skew Ratio (DSR) trends for 10 classes showing critical points at different missing rates. K_{crit} decreases with increasing missing rate (0.3, 0.5, 0.7). (b) DSR trends for 20 classes. Compared to 10 classes, K_{crit} shifts rightward, confirming the proportional relationship between K_{crit} and C.

5.5 Comprehensive Performance Evaluation

Our theoretical framework is robustly supported by consistent divergence trends observed across all evaluation metrics under various client counts and missing rates, including average accuracy, loss, accuracy variance, and class-wise performance gaps; detailed supporting analyses are provided in Appendix C.3, particularly Sections C.3.1–C.3.3, and Appendix C.4.

5.6 SSDI-Guided Algorithm Selection

Beyond characterizing heterogeneity, we investigate whether SSDI can guide federated algorithm selection across heterogeneity regimes.

Experimental Setup. We compare four representative federated learning algorithms: FedAvg, FedProx, SCAFFOLD, and Ditto on MNIST with $C = 10$. Datasets are generated with four SSDI levels $(0.2, 0.4, 0.6, 0.8)$ under two client scales ($K = 20$ and $K = 300$). To reduce randomness caused by data partitioning, each experiment is conducted on five independently generated distribution matrices with different seeds. We report the mean and standard deviation across these runs.

The complete numerical results are provided in Appendix C.6, including average test accuracy (Table 12), worst-class accuracy (Table 13), and the standard deviation of class-wise accuracy (Table 14).

Analysis. Figure 5 reveals three distinct regimes across SSDI levels; full numerical results are provided in Appendix C.6.

When SSDI is low (≈ 0.2), all algorithms achieve very similar performance, indicating that heterogeneity is mild and does not significantly affect optimization dynamics.

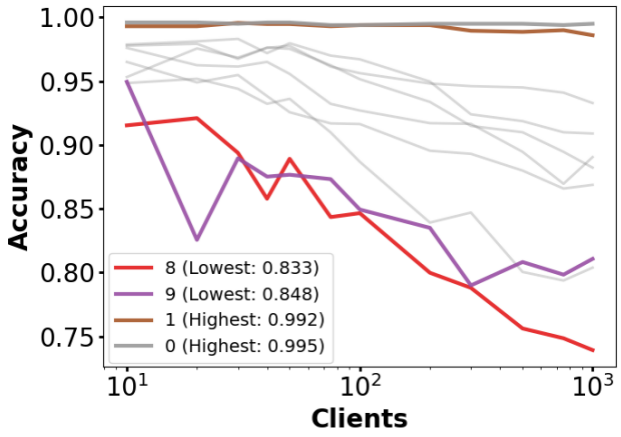
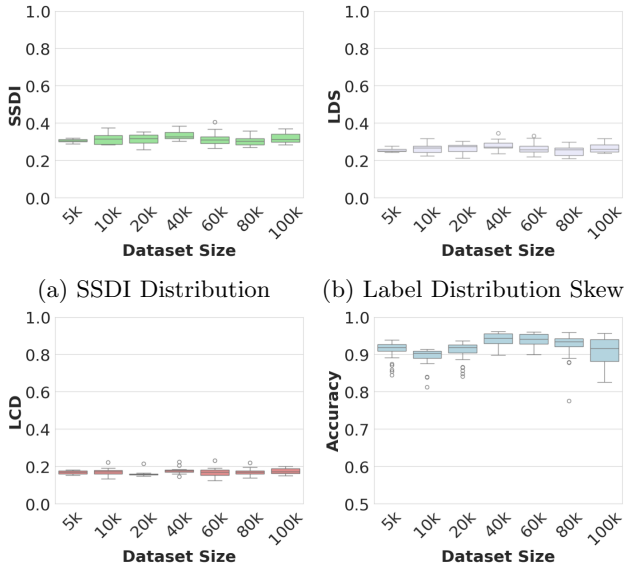


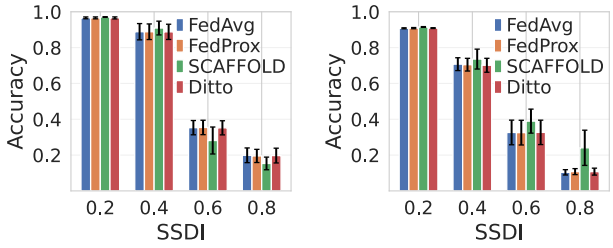
Figure 3: Class-wise performance analysis for 10 classes with missing rate 0.3. The accuracy trends reveal increasing performance divergence: rare label accuracy deteriorates significantly with increasing clients, while frequent labels maintain high accuracy.



(a) SSDI Distribution (b) Label Distribution Skew

(c) Label Coverage Deficiency (d) Accuracy

Figure 4: Robustness analysis of SSDI components across different dataset sizes: SSDI, LDS, LCD and corresponding accuracy trends.



(a) $K = 20$ (b) $K = 300$

Figure 5: Algorithm performance across SSDI regimes for different client scales.

Under moderate SSDI (≈ 0.4), SCAFFOLD consistently outperforms other methods, especially when the number of clients is large. This behavior aligns with our theoretical interpretation that heterogeneity in this regime is dominated by Label Distribution Skew (LDS), which introduces client drift during local training. SCAFFOLD explicitly corrects such drift using control variates, leading to improved convergence and higher final accuracy.

When SSDI becomes high (≥ 0.6), overall performance deteriorates sharply across all algorithms. Although SCAFFOLD still shows relatively better results in some settings, the accuracy of all methods remains low due to severe Label Coverage Deficiency (LCD), where certain classes become extremely rare or completely missing across many clients. In this regime, the fundamental limitation is no longer optimization bias but the lack of sufficient label coverage in the distributed data. Consequently, algorithmic improvements alone cannot fully mitigate the performance degradation, and improving label coverage becomes essential.

These results demonstrate that SSDI not only quantifies the severity of data heterogeneity but also provides diagnostic guidance for selecting suitable federated optimization strategies under different heterogeneity regimes.

6 CONCLUSION

We have introduced the Skew-Scarcity Disentanglement Indicator (SSDI) to disentangle federated learning heterogeneity into the orthogonal Label Distribution Skew (LDS) and Label Coverage Deficiency (LCD).

Our analysis suggests that, as the number of clients increases, system behaviour can shift around a critical point $K_{\text{crit}} \propto C \ln C$, where the dominant heterogeneity mechanism transitions from distribution skew to coverage deficiency. Under the studied regimes, this transition can lead to catastrophic accuracy collapse on rare labels. Crucially, our theoretical and empirical results suggest that overall dataset size has limited effect on this transition, while higher missing rates not only intensify performance divergence but also reduce K_{crit}, making systems more vulnerable to heterogeneity risks. Experimental validation confirms SSDI's accuracy in capturing these transitions, providing a fine-grained diagnostic framework for developing targeted mitigation strategies in federated systems.

References

McMahan B., Moore E., Ramage D., Hampson S., & Arcas B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273-1282).

Li T., Sahu A. K., Zaheer M., Sanjabi M., Talwalkar A., & Smith V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429-450.

Konečný J., McMahan H. B., Ramage D., & Richtárik P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Li T., Sahu A. K., Talwalkar A., & Smith V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3), 50-60.

Zhu H., Xu J., Liu S., & Jin Y. (2021). Federated learning on non-IID data: A survey. *Neurocomputing*, 465, 371-390.

Lu Z., Pan H., Dai Y., Si X., & Zhang Y. (2024). Federated learning with non-IID data: A survey. *IEEE Internet of Things Journal*, 11(11), 19188-19209.

Shuai X., Shen Y., Jiang S., Zhao Z., Yan Z., & Xing G. (2022). BalanceFL: Addressing class imbalance in long-tail federated learning. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)* (pp. 271-284).

McAllester D. A. (1999). PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 164-170).

Domini, D., Ingemann, C. O., Aguzzi, G., Esterle, L., & Viroli, M. (2026). ProFed: A benchmark for proximity-based non-IID federated learning. *Journal of Open Research Software*, 14, 13. doi:10.5334/jors.624.

Hsu T. M. H., Qi H., & Brown M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Yang C., Wang Q., Xu M., Chen Z., Bian K., Liu Y., & Liu X. (2021). Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *Proceedings of the Web Conference 2021* (pp. 935-946).

Shi, Y., Liang, J., Zhang, W., Xue, C., Tan, V. Y. F., & Bai, S. (2024). Understanding and mitigating dimensional collapse in federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5), 2936–2949. doi:10.1109/TPAMI.2023.3338063.

Karimireddy S. P., Kale S., Mohri M., Reddi S., Stich S., & Suresh A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning* (pp. 5132-5143).

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, H. B. (2021). Adaptive federated optimization. *International Conference on Learning Representations (ICLR)*.

Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2020). On the convergence of FedAvg on non-IID data. *International Conference on Learning Representations (ICLR)*.

Sun, Y., Shen, L., & Tao, D. (2026). Towards understanding generalization and stability gaps between centralized and decentralized federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, Y., Shen, L., & Tao, D. (2023). Which mode is better for federated learning? Centralized or decentralized. *ResearchGate preprint*. doi:10.13140/RG.2.2.22032.79362.

He F., Liu T., & Tao D. (2019). Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems* (Vol. 32).

Hardt M., Recht B., & Singer Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning* (pp. 1225-1234).

Li T., Hu S., Beirami A., & Smith V. (2021). Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning* (pp. 6357-6368).

T Dinh C., Tran N., & Nguyen J. (2020). Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 21394-21405).

Yi L., Yu H., Ren C., Zhang H., Wang G., Liu X., & Li X. (2024). pFedAFM: Adaptive feature mixture for batch-level personalization in heterogeneous federated learning. *arXiv preprint arXiv:2404.17847*.

Yang N., Chen X., Liu C. Z., Yuan D., Bao W., & Cui L. (2023). FedMAE: Federated self-supervised learning with one-block masked auto-encoder. *arXiv preprint arXiv:2303.11339*.

Wang, L., Zhang, K., Li, Y., Tian, Y., & Tedrake, R. (2023). Does learning from decentralized non-IID unlabeled data benefit from self supervision? *International Conference on Learning Representations (ICLR)*.

Iakovidou C., & Kim K. (2024). Asynchronous federated stochastic optimization for heterogeneous objectives under arbitrary delays. *arXiv preprint arXiv:2405.10123*.

Kang Y., & Li B. (2024). Polaris: Accelerating asynchronous federated learning with client selection. *IEEE Transactions on Cloud Computing*, 12(2), 446-458.

Chen, H.-Y., & Chao, W.-L. (2021). FedBE: Making Bayesian model ensemble applicable to federated learning. *International Conference on Learning Representations (ICLR)*.

Lecun Y., Bottou L., Bengio Y., & Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. doi:10.1109/5.726791

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical Report*, University of Toronto.

Deng J., Dong W., Socher R., Li L.-J., Li K., & Li F.-F. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. doi:10.1109/CVPR.2009.5206848

He K., Zhang X., Ren S., & Sun J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma N., Zhang X., Zheng H.-T., & Sun J. (2018). ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes]
   Justification: Section 3 and Section 4 provide detailed descriptions of the SSDI metric and the PAC-Bayesian framework.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes]
   Justification: The computational and memory complexity of SSDI are analyzed in Appendix C.4.2 and summarized in Appendix C.5, especially Algorithm 2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [No]
   Justification: We do not include anonymized source code directly in the manuscript. Code and implementation details are instead provided through the associated repository.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes]
   Justification: Section 4.1 states the core model assumptions, and further justification and theoretical motivation are provided in Appendix B.6 and Appendix B.8.
   (b) Complete proofs of all theoretical results.
   [Yes]
   Justification: Proofs of Theorem 1, Theorem 2, Corollary 1, Corollary 2, Corollary 3, Corollary 4, and Proposition 1 are provided in Appendix B.1, Appendix B.3, Appendix B.5, Appendix B.6, and Appendix B.7.
   (c) Clear explanations of any assumptions.
   [Yes]
   Justification: Assumptions are explained in Section 4.1, with detailed physical interpretations and theoretical motivation provided in Appendix B.6 and Appendix B.8.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes]
   Justification: Code and implementation details are provided in the associated repository, with reproducibility guidelines in Appendix C.5.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes]
   Justification: Section 5.1 describes the datasets, models, and non-IID partitioning strategy, with full experimental details provided in Appendix C.1.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Yes]

Justification: Evaluation metrics are defined in Section 5.1, and additional empirical analyses are provided in Appendix C.2, Appendix C.3, Appendix C.4, and Appendix C.6.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
[Yes]
Justification: Computational infrastructure is described in Appendix C.5, especially Appendix C.5.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets.
[Yes]
Justification: Datasets like MNIST, CIFAR-100, and Tiny ImageNet are standard and properly cited in the references.

(b) The license information of the assets, if applicable.
[No]
Justification: License information is not explicitly mentioned, but datasets are publicly available for research use.

(c) New assets either in the supplemental material or as a URL, if applicable.
[Yes]
Justification: New code is provided in the associated repository, with algorithmic and implementation details in Appendix C.5.

(d) Information about consent from data providers/curators.
[Not Applicable]
Justification: All datasets are publicly available and do not require additional consent.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
[Not Applicable]
Justification: The datasets used do not contain sensitive content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
[Not Applicable]
Justification: No human subjects were involved.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
[Not Applicable]
Justification: No human subjects were involved.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
[Not Applicable]
Justification: No crowdsourcing was used.

Supplementary Material for Disentangling Federated Learning Heterogeneity

A Notation and Symbol Summary

Table 2: Detailed summary of notation used throughout the paper.

Symbol	Meaning
K	Number of clients
C	Number of classes
N_{total}	Total number of samples in the federated system
$n_{k,c}$	Number of samples of class c on client k
n_k	Total number of samples on client k, defined as $n_k = \sum_{c=1}^{C} n_{k,c}$
N_c	Total number of samples of class c, defined as $N_c = \sum_{k=1}^{K} n_{k,c}$
$\rho_{k,c}$	Local class proportion of class c on client k, defined as $\rho_{k,c} = n_{k,c}/n_k$
ρ_c^{global}	Global class proportion of class c, defined as $\rho_c^{\text{global}} = N_c/N_{\text{total}}$
d^c	Class-level LCD deviation associated with class c
d_k	Client-level LDS deviation associated with client k
$\mathbf{v}$	Unified deviation vector formed by concatenating $\{d^c\}_{c=1}^{C}$ and $\{d_k\}_{k=1}^{K}$
P	Empirical joint distribution matrix in $\mathbb{R}^{C\times K}$
$p_{c,k}$	Entry of P, defined as $p_{c,k} = n_{k,c}/N_{\text{total}}$
Q	Ideal independent distribution matrix in $\mathbb{R}^{C\times K}$
$q_{c,k}$	Entry of Q, defined as $q_{c,k} = \rho_c^{\text{global}} \cdot n_k/N_{\text{total}}$
D	Deviation matrix, defined as $D = P - Q$, measuring departure from the ideal IID case
$d_{c,k}$	Entry of D, defined as $d_{c,k} = p_{c,k} - q_{c,k}$
W	Weight matrix used in the weighted Frobenius norm
$w_{c,k}$	Entry of W, defined as $w_{c,k} = \sqrt{N_{\text{total}}/N_c + N_{\text{total}}/n_k}$
$\|\mathbf{v}_{\max}\|_2$	(C,K)-adaptive theoretical normalization constant, defined as $\sqrt{2\,(1 - 1/\min(C,K))}$
SSDI	Overall Skew-Scarcity Disentanglement Indicator
SSDI_{LDS}	Label Distribution Skew component of SSDI
SSDI_{LCD}	Label Coverage Deficiency component of SSDI
DSR	Deficiency-to-Skew Ratio, defined as $\text{SSDI}_{\text{LCD}}/\text{SSDI}_{\text{LDS}}$
$P(\theta)$	Prior distribution over model parameters
$Q(\theta)$	Posterior distribution over model parameters
μ	Posterior mean vector; $\|\mu\|_2^2$ measures overall mean drift
μ_k^2	Effective client-level drift magnitude in the penalty decomposition
μ_c^2	Effective class-level drift magnitude in the penalty decomposition
Σ_q	Posterior covariance matrix
$\sigma_{q,i}^2$	Posterior variance along parameter dimension i
$\sigma_{q,k}^2$	Effective client-level uncertainty term in the penalty decomposition
$\sigma_{q,c}^2$	Effective class-level uncertainty term in the penalty decomposition
Σ_p	Prior covariance matrix
γ	Global regularization strength coefficient in the adaptive prior
ϕ(SSDI)	Heterogeneity strength function in the adaptive prior
ω_0	Baseline weight vector in the prior precision matrix
Γ_{LDS}	LDS-related penalty term in the generalization bound
Γ_{LCD}	LCD-related penalty term in the generalization bound
Ψ	Lower-order model complexity term
$M_{c,k}$	Label missingness indicator, equal to 1 if $n_{k,c} = 0$ and 0 otherwise
K_{crit}	Critical client scale at which dominance shifts between LDS and LCD
κ	System-dependent constant in the scaling law $K_{\text{crit}} \propto \kappa C \ln C$

B Theoretical Proofs and Supporting Derivations

The supplementary materials contain detailed proofs of results omitted from the main paper.

B.1 Equivalent Representations and Basic Properties of SSDI

B.1.1 Equivalence Between the Vector and Matrix Forms of SSDI

We prove the equivalence between the vector form and matrix form representations of SSDI.

Proof. Starting from the vector definition in the main paper

$$\|\mathbf{v}\|_2^2 = \sum_{c=1}^{C} (d^c)^2 + \sum_{k=1}^{K} (d_k)^2 . \tag{17}$$

Recall the deviation definitions from Section 3.2 of the main paper

$$d^c = \sqrt{\frac{1}{N_c} \cdot \frac{1}{N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}, \tag{18}$$

$$d_k = \sqrt{\frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}. \tag{19}$$

We first expand the squares of the deviation terms

$$(d^c)^2 = \frac{1}{N_c} \cdot \frac{1}{N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2, \tag{20}$$

$$(d_k)^2 = \frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2 . \tag{21}$$

Now, express the squared deviations in terms of P and Q. Recall that

$$\rho_{k,c} - \rho_c^{\text{global}} = \frac{N_{\text{total}}}{n_k} \left(\frac{n_{k,c}}{N_{\text{total}}} - \rho_c^{\text{global}} \cdot \frac{n_k}{N_{\text{total}}} \right) = \frac{N_{\text{total}}}{n_k} \left(p_{c,k} - q_{c,k}\right). \tag{22}$$

Substituting into

$$\begin{aligned} (d^c)^2 &= \frac{1}{N_c} \cdot \frac{1}{N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \left[\frac{N_{\text{total}}}{n_k} \left(p_{c,k} - q_{c,k}\right) \right]^2 = \frac{1}{N_c} \cdot \frac{1}{N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \cdot \frac{N_{\text{total}}^2}{n_k^2} \left(p_{c,k} - q_{c,k}\right)^2 \\ &= \frac{1}{N_c} \sum_{k=1}^{K} N_{\text{total}} \left(p_{c,k} - q_{c,k}\right)^2 . \end{aligned} \tag{23}$$

Substituting into

$$\begin{aligned} (d_k)^2 &= \frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \left[\frac{N_{\text{total}}}{n_k} \left(p_{c,k} - q_{c,k}\right) \right]^2 = \frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \frac{N_{\text{total}}^2}{n_k^2} \left(p_{c,k} - q_{c,k}\right)^2 \\ &= \frac{N_{\text{total}}}{n_k} \sum_{c=1}^{C} \left(p_{c,k} - q_{c,k}\right)^2 . \end{aligned} \tag{24}$$

Now, sum these terms over all classes and clients:

$$\sum_{c=1}^{C}(d^c)^2=\sum_{c=1}^{C}\left[\frac{1}{N_c}\sum_{k=1}^{K}N_{\text{total}}\left(p_{c,k}-q_{c,k}\right)^2\right]=\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{N_{\text{total}}}{N_c}\left(p_{c,k}-q_{c,k}\right)^2, \tag{25}$$

$$\sum_{k=1}^{K}(d_k)^2=\sum_{k=1}^{K}\left[\frac{N_{\text{total}}}{n_k}\sum_{c=1}^{C}\left(p_{c,k}-q_{c,k}\right)^2\right]=\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{N_{\text{total}}}{n_k}\left(p_{c,k}-q_{c,k}\right)^2. \tag{26}$$

Combining both sums

$$\|\mathbf{v}\|_2^2=\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{N_{\text{total}}}{N_c}\left(p_{c,k}-q_{c,k}\right)^2+\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{N_{\text{total}}}{n_k}\left(p_{c,k}-q_{c,k}\right)^2=\sum_{c=1}^{C}\sum_{k=1}^{K}\left(\frac{N_{\text{total}}}{N_c}+\frac{N_{\text{total}}}{n_k}\right)\left(p_{c,k}-q_{c,k}\right)^2. \tag{27}$$

Define the weight matrix W with elements

$$w_{c,k}=\sqrt{\frac{N_{\text{total}}}{N_c}+\frac{N_{\text{total}}}{n_k}}, \tag{28}$$

then we have

$$\|\mathbf{v}\|_2^2=\sum_{c=1}^{C}\sum_{k=1}^{K}w_{c,k}^2(p_{c,k}-q_{c,k})^2=\|W\odot D\|_F^2. \tag{29}$$

For the maximum deviation case (extreme heterogeneity), we have

$$\|\mathbf{v}_{\max}\|_2=\|W_{\max}\odot D_{\max}\|_F. \tag{30}$$

Therefore, the two representations are equivalent

$$\text{SSDI}=\frac{\|\mathbf{v}\|_2}{\|\mathbf{v}_{\max}\|_2}=\frac{\|W\odot D\|_F}{\|W_{\max}\odot D_{\max}\|_F}. \tag{31}$$

This establishes the equivalence between the vector and matrix representations of SSDI. □

B.1.2 Properties of Proposition 2

Proposition 2. *The SSDI metric satisfies* $0\le SSDI\le 1$.

Proof. For the lower bound, by non-negativity of the Frobenius norm,

$$\|W\odot D\|_F\ge 0, \tag{32}$$

and the normalization constant $V_{\max}(C,K)$ is strictly positive whenever $\min(C,K)\ge 2$. Hence SSDI ≥ 0. Equality holds when $D=0$, i.e., when $P=Q$, corresponding to the perfectly IID case.

For the upper bound, by construction,

$$V_{\max}(C,K)=\sup_{P\in\mathcal{P}_{C,K}}\|W\odot D\|_F. \tag{33}$$

Therefore, for any admissible joint distribution P,

$$\|W\odot D\|_F\le V_{\max}(C,K), \tag{34}$$

which implies

$$\text{SSDI}=\frac{\|W\odot D\|_F}{V_{\max}(C,K)}\le 1. \tag{35}$$

Under the theoretical convention allowing zero-mass classes or clients in the limiting extremal construction, this upper bound is attained in the limiting sense by the extremal matching skeleton described in Appendix B.2. □

B.2 Definitions of SSDI-Related Matrices

This section provides formal definitions and key properties of the matrices that form the foundation of the SSDI metric.

B.2.1 Empirical Joint Distribution Matrix P

Definition: $P \in \mathbb{R}^{C \times K}$, defined as $p_{c,k} = \frac{n_{k,c}}{N_{\text{total}}}$.

Properties:

- **Non-negativity:** $p_{c,k} \geq 0$ for all c, k.
- **Normalisation:** $\sum_{c=1}^{C} \sum_{k=1}^{K} p_{c,k} = 1$.
- **Marginal Distributions:** $\sum_{k=1}^{K} p_{c,k} = \frac{N_c}{N_{\text{total}}}$ and $\sum_{c=1}^{C} p_{c,k} = \frac{n_k}{N_{\text{total}}}$.

B.2.2 Ideal Independent Distribution Matrix Q

Definition: $Q \in \mathbb{R}^{C \times K}$, defined as $q_{c,k} = \rho_c^{\text{global}} \cdot \frac{n_k}{N_{\text{total}}} = \frac{N_c}{N_{\text{total}}} \cdot \frac{n_k}{N_{\text{total}}}$.

Properties:

- **Independence Assumption:** Represents the ideal case where client and label distributions are independent.
- **Same Marginals as P:** The row and column sums of Q match the marginal distributions of P.
- **normalisation:** $\sum_{c,k} q_{c,k} = 1$.

B.2.3 Deviation Matrix D

Definition: $D = P - Q$.

Properties:

- **Zero-Sum:** $\sum_{c,k} d_{c,k} = 0$.
- **Marginal Zero-Sum:** $\sum_k d_{c,k} = 0$ and $\sum_c d_{c,k} = 0$ for all c and k.
- **Statistical Meaning:** Quantifies the statistical dependence between clients and labels.

B.2.4 Maximum Deviation and Weight Matrices

Let

$$m = \min(C, K). \tag{36}$$

We define the normalization through the theoretical supremum over the unconstrained joint-distribution space, where some classes or clients are allowed to approach zero mass in the limiting extremal configuration. When $C = K$, the limiting extremal structure reduces to a permutation-type extreme structure. When $C \neq K$, it is governed by a rectangular matching skeleton of effective size m, while the remaining rows or columns vanish in the limiting sense.

After re-indexing the active rows and columns, a canonical representative of the limiting extremal structure is

$$(P_{\max})_{c,k} = \begin{cases} \frac{1}{m}, & c = k \leq m, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

Its induced independent counterpart is

$$(Q_{\max})_{c,k} = \begin{cases} \frac{1}{m^2}, & c \leq m,\ k \leq m, \\ 0, & \text{otherwise,} \end{cases} \tag{38}$$

and thus the corresponding deviation matrix is

$$(D_{\max})_{c,k} = \begin{cases} \frac{m-1}{m^2}, & c = k \le m, \\ -\frac{1}{m^2}, & c \ne k,\ c \le m,\ k \le m, \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

For each active class and active client, the marginal masses equal $1/m$. Therefore, on the active $m \times m$ block,

$$(W_{\max})_{c,k} = \sqrt{\frac{1}{1/m} + \frac{1}{1/m}} = \sqrt{2m}. \tag{40}$$

The remaining rows or columns correspond to vanishing class or client mass in the limiting construction. Hence $W_{\max}$ should be understood as an active-block limiting quantity, while the product $W_{\max} \odot D_{\max}$ remains well-defined in the normalization analysis.

Accordingly, the (C, K)-adaptive normalization constant is

$$\|\mathbf{v}_{\max}\|_2 = V_{\max}(C, K) = \sqrt{2\left(1 - \frac{1}{\min(C, K)}\right)}. \tag{41}$$

Remark on Practical Feasibility: If one additionally requires every class and every client to be strictly non-empty, then the above supremum may become unattainable when $C \ne K$, although it remains the correct theoretical normalization reference.

B.2.5 Proof of the Normalized Range of SSDI

To justify that SSDI $\in [0, 1]$ under the (C, K)-adaptive normalization, it suffices to evaluate the deviation norm induced by the limiting extremal matrices defined in Section B.2.4.

Let

$$m = \min(C, K). \tag{42}$$

Using the explicit form of $D_{\max}$ together with the active-block weight $(W_{\max})_{c,k} = \sqrt{2m}$, we obtain

$$\|W_{\max} \odot D_{\max}\|_F^2 = \sum_{c=1}^{m} \sum_{k=1}^{m} \left(\sqrt{2m}\,(D_{\max})_{c,k}\right)^2. \tag{43}$$

Separating diagonal and off-diagonal terms yields

$$\|W_{\max} \odot D_{\max}\|_F^2 = 2m \left[m \left(\frac{m-1}{m^2}\right)^2 + m(m-1)\left(\frac{1}{m^2}\right)^2 \right]. \tag{44}$$

Simplifying gives

$$\|W_{\max} \odot D_{\max}\|_F^2 = 2m \left[\frac{(m-1)^2}{m^3} + \frac{m-1}{m^3}\right] = 2m \cdot \frac{m-1}{m^2} = 2\left(1 - \frac{1}{m}\right). \tag{45}$$

Hence,

$$\sup_{P \in \mathcal{P}_{C,K}} \|\mathbf{v}(P)\|_2 = \sup_{P \in \mathcal{P}_{C,K}} \|W \odot D\|_F = \sqrt{2\left(1 - \frac{1}{m}\right)} = \sqrt{2\left(1 - \frac{1}{\min(C, K)}\right)}. \tag{46}$$

If we define the normalization constant as

$$\|\mathbf{v}_{\max}\|_2 = V_{\max}(C, K) = \sqrt{2\left(1 - \frac{1}{\min(C, K)}\right)}, \tag{47}$$

then it follows that

$$0 \leq \mathrm{SSDI} = \frac{\|\mathbf{v}\|_2}{\|\mathbf{v}_{\max}\|_2} \leq 1. \tag{48}$$

This establishes the normalized range of SSDI under the theoretical convention that zero-mass classes or zero-mass clients are allowed in the limiting extremal construction.

Experimental non-empty convention: In our experiments, we retain the practically meaningful requirement that every class and every client be non-empty. Under this additional feasibility constraint, the above theoretical supremum can become unattainable when $C \neq K$, so the empirical maximum SSDI may be strictly smaller than 1. Nevertheless, the theoretical $V_{\max}(C, K)$ remains a stable normalization reference across different (C, K) settings.

B.3 Orthogonal Decomposition of SSDI

B.3.1 Lemma 1: LDS–LCD orthogonality

Lemma 1. *The LDS and LCD components are orthogonal under the weighted Frobenius inner product:*

$$\langle W \odot D^{LDS}, W \odot D^{LCD} \rangle_F = 0. \tag{49}$$

Proof. Define the presence and missingness deviation matrices

$$D^{\mathrm{LDS}}_{c,k} = \begin{cases} p_{c,k} - q_{c,k}, & n_{k,c} > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

$$D^{\mathrm{LCD}}_{c,k} = \begin{cases} 0, & n_{k,c} > 0; \\ -q_{c,k}, & n_{k,c} = 0. \end{cases} \tag{51}$$

By construction, for any client-class pair (c, k), at least one of $D^{\mathrm{LDS}}_{c,k}$ and $D^{\mathrm{LCD}}_{c,k}$ is zero. Therefore,

$$D^{\mathrm{LDS}}_{c,k} \cdot D^{\mathrm{LCD}}_{c,k} = 0, \quad \text{for all } (c, k). \tag{52}$$

The weighted Frobenius inner product expands as

$$\langle W \odot D^{\mathrm{LDS}}, W \odot D^{\mathrm{LCD}} \rangle_F = \sum_{c=1}^{C} \sum_{k=1}^{K} (w_{c,k} D^{\mathrm{LDS}}_{c,k})(w_{c,k} D^{\mathrm{LCD}}_{c,k}). \tag{53}$$

This simplifies to

$$\langle W \odot D^{\mathrm{LDS}}, W \odot D^{\mathrm{LCD}} \rangle_F = \sum_{c=1}^{C} \sum_{k=1}^{K} W^2_{c,k} (D^{\mathrm{LDS}}_{c,k} D^{\mathrm{LCD}}_{c,k}). \tag{54}$$

Since $D^{\mathrm{LDS}}_{c,k} D^{\mathrm{LCD}}_{c,k} = 0$ for all (c, k), we have

$$\langle W \odot D^{\mathrm{LDS}}, W \odot D^{\mathrm{LCD}} \rangle_F = 0. \tag{55}$$

This completes the proof of orthogonality between LDS and LCD components under the weighted Frobenius inner product. □

B.3.2 Proof of Proposition 3

Proposition 3. *SSDI can be decomposed into the square root of the sum of squares of LDS and LCD components:*

$$SSDI = \sqrt{(SSDI_{LDS})^2 + (SSDI_{LCD})^2}. \tag{56}$$

Proof. By definition, the total deviation matrix decomposes as

$$D = D^{\mathrm{LDS}} + D^{\mathrm{LCD}}. \tag{57}$$

The squared weighted Frobenius norm of the total deviation is:

$$\|W \odot D\|_F^2 = \|W \odot (D^{\mathrm{LDS}} + D^{\mathrm{LCD}})\|_F^2. \tag{58}$$

Expanding this expression

$$\|W \odot D\|_F^2 = \|W \odot D^{\mathrm{LDS}}\|_F^2 + \|W \odot D^{\mathrm{LCD}}\|_F^2 + 2\langle W \odot D^{\mathrm{LDS}}, W \odot D^{\mathrm{LCD}}\rangle_F. \tag{59}$$

From Lemma 1, we know

$$\langle W \odot D^{\mathrm{LDS}}, W \odot D^{\mathrm{LCD}}\rangle_F = 0. \tag{60}$$

Therefore, the cross-term vanishes

$$\|W \odot D\|_F^2 = \|W \odot D^{\mathrm{LDS}}\|_F^2 + \|W \odot D^{\mathrm{LCD}}\|_F^2. \tag{61}$$

Recall the SSDI component definitions

$$\mathrm{SSDI}_{\mathrm{LDS}} = \frac{\|W \odot D^{\mathrm{LDS}}\|_F}{\|W_{\max} \odot D_{\max}\|_F}, \tag{62}$$

$$\mathrm{SSDI}_{\mathrm{LCD}} = \frac{\|W \odot D^{\mathrm{LCD}}\|_F}{\|W_{\max} \odot D_{\max}\|_F}, \tag{63}$$

$$\mathrm{SSDI} = \frac{\|W \odot D\|_F}{\|W_{\max} \odot D_{\max}\|_F}. \tag{64}$$

Substituting the norm decomposition

$$\mathrm{SSDI}^2 = \left(\frac{\|W \odot D\|_F}{\|W_{\max} \odot D_{\max}\|_F}\right)^2 = \frac{\|W \odot D^{\mathrm{LDS}}\|_F^2 + \|W \odot D^{\mathrm{LCD}}\|_F^2}{\|W_{\max} \odot D_{\max}\|_F^2}. \tag{65}$$

This simplifies to

$$\mathrm{SSDI}^2 = \left(\frac{\|W \odot D^{\mathrm{LDS}}\|_F}{\|W_{\max} \odot D_{\max}\|_F}\right)^2 + \left(\frac{\|W \odot D^{\mathrm{LCD}}\|_F}{\|W_{\max} \odot D_{\max}\|_F}\right)^2 = (\mathrm{SSDI}_{\mathrm{LDS}})^2 + (\mathrm{SSDI}_{\mathrm{LCD}})^2. \tag{66}$$

Taking the square root of both sides completes the proof:

$$\mathrm{SSDI} = \sqrt{(\mathrm{SSDI}_{\mathrm{LDS}})^2 + (\mathrm{SSDI}_{\mathrm{LCD}})^2}. \tag{67}$$

□

B.4 Fine-Grained Two-Dimensional Decomposition of SSDI

To improve interpretability and provide finer diagnostic insight, we further refine the SSDI components along two orthogonal axes: the *class axis* and the *client axis*. This two-dimensional refinement provides localized diagnostic summaries of Label Distribution Skew (LDS) and Label Coverage Deficiency (LCD) at different granularities, without introducing additional assumptions beyond the main formulation.

Fine-grained deviation metrics: Starting from the SSDI decomposition, we define four deviation metrics that characterize heterogeneity at different levels:

Class-level Coverage Deviation $d_c^{\mathbf{LCD}}$ identifies classes that suffer from severe under-coverage or complete absence across clients.

Client-level Coverage Deviation $d_k^{\mathbf{LCD}}$ identifies clients with insufficient label diversity due to missing classes.

Class-level Distribution Deviation $d_c^{\mathbf{LDS}}$ captures uneven distributions of class frequencies across clients where the label is present.

Client-level Distribution Deviation $d_k^{\mathbf{LDS}}$ captures skewed local label distributions within individual clients.

Together, these deviations help identify whether heterogeneity originates primarily from specific classes, specific clients, or their interaction.

Aggregation-level diagnosis: To summarize heterogeneity patterns at the system level, we aggregate the fine-grained deviations along each axis. This yields two interpretable diagnostic indicators for each SSDI component:

A high $H_{\mathbf{coverage}}$ indicates a system-wide *Label Coverage Deficiency*, suggesting mitigation strategies such as data augmentation, coverage-aware client selection, or personalized modelling.

A high $H_{\mathbf{distribution}}$ indicates a system-wide *Label Distribution Skew*, suggesting strategies such as loss re-weighting or client drift correction.

This hierarchical structure allows practitioners to move from a global heterogeneity score (SSDI) to targeted diagnosis along the class or client dimension.

Two-dimensional interpretation. The resulting refinement organizes heterogeneity into four interpretable diagnostic components, summarized in Table 3.

Table 3: Two-dimensional decomposition of federated learning heterogeneity.

	$H_{\mathbf{coverage}}$ **(Class-axis)**	$H_{\mathbf{distribution}}$ **(Client-axis)**
LCD Component	$H^{\mathrm{LCD}}_{\mathrm{coverage}}$: Class coverage problem	$H^{\mathrm{LCD}}_{\mathrm{distribution}}$: Client missingness problem
LDS Component	$H^{\mathrm{LDS}}_{\mathrm{coverage}}$: Class distribution problem	$H^{\mathrm{LDS}}_{\mathrm{distribution}}$: Client skew problem

Conceptual mapping to classical heterogeneity: From a conceptual perspective, **label shift** is primarily reflected in the LDS row, where $H^{\mathrm{LDS}}_{\mathrm{distribution}}$ captures client-specific distribution skew and $H^{\mathrm{LDS}}_{\mathrm{coverage}}$ captures class-level imbalance among *present* labels. In contrast, **quantity shift** is primarily reflected in the LCD row, where $H^{\mathrm{LCD}}_{\mathrm{coverage}}$ measures class-level coverage deficiency (including complete absence), and $H^{\mathrm{LCD}}_{\mathrm{distribution}}$ captures client-level missingness patterns.

Mathematical definition: The aggregated diagnostic indicators are defined as normalized ℓ_2 norms:

$$H^{*}_{\mathrm{coverage}} = \frac{\|\mathbf{d}^{c}_{*}\|_2}{\|\mathbf{v}_{\max}\|_2}, \quad H^{*}_{\mathrm{distribution}} = \frac{\|\mathbf{d}^{k}_{*}\|_2}{\|\mathbf{v}_{\max}\|_2},$$

where $* \in \{\mathrm{LCD}, \mathrm{LDS}\}$, $\mathbf{d}^{c}_{*} = (d^{*}_1, \ldots, d^{*}_C)$ denotes deviations along the class axis, $\mathbf{d}^{k}_{*} = (d^{*}_1, \ldots, d^{*}_K)$ denotes deviations along the client axis, and $\|\mathbf{v}_{\max}\|_2$ is the normalization constant defined in the main paper.

Overall, this fine-grained refinement enhances the interpretability of SSDI while remaining fully consistent with the main formulation. It provides an auxiliary diagnostic tool for identifying whether heterogeneity is dominated by coverage deficiency or distribution skew, and whether it manifests primarily at the class or client level.

B.5 Proof of Theorem 1

B.5.1 Main Proof Structure

Proof. We begin with the standard PAC-Bayes inequality for any prior distribution P independent of the training data:

$$R(Q) \leq \hat{R}(Q) + \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \log\frac{1}{\delta}}{2N_{\mathrm{total}} - 1}}. \tag{68}$$

We design an adaptive prior distribution $P(\boldsymbol{\theta}) = \mathcal{N}(0, \boldsymbol{\Sigma}_p)$ with

$$\boldsymbol{\Sigma}_p^{-1} = \gamma \cdot \mathrm{Diag}(\boldsymbol{\omega}), \quad \boldsymbol{\omega} = \phi(\mathrm{SSDI}) \cdot \boldsymbol{\omega}_0, \tag{69}$$

where

$$\phi(\mathrm{SSDI}) = \frac{\mathrm{SSDI}}{1 - \mathrm{SSDI} + \epsilon}, \tag{70}$$

$$\boldsymbol{\omega}_0 = \left(\frac{N_{\mathrm{total}}}{N_1}, \ldots, \frac{N_{\mathrm{total}}}{N_C}, \frac{N_{\mathrm{total}}}{n_1}, \ldots, \frac{N_{\mathrm{total}}}{n_K}\right)^T. \tag{71}$$

The posterior distribution $Q(\boldsymbol{\theta}) = \mathcal{N}(\mu, \boldsymbol{\Sigma}_q)$ with $\boldsymbol{\Sigma}_q = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$ satisfies the parameter drift constraint:

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}\|^2 \leq \frac{2}{\lambda_{\mathrm{strong}}} D_{\mathrm{KL}}(Q\|P). \tag{72}$$

From Lemma 2, the KL divergence between posterior and prior is

$$D_{\mathrm{KL}}(Q\|P) = \frac{1}{2}\left[\gamma\phi(\mathrm{SSDI}) \sum_{i=1}^{d} \boldsymbol{\omega}_{0,i}(\sigma_i^2 + \mu_i^2) - d\log\phi(\mathrm{SSDI}) - \sum_{i=1}^{d} \log(\gamma\boldsymbol{\omega}_{0,i}\sigma_i^2) - d\right]. \tag{73}$$

Separating the KL divergence into heterogeneity-dependent and complexity terms

$$D_{\mathrm{KL}}(Q\|P) = \frac{1}{2}\left[\gamma\phi(\mathrm{SSDI}) \sum_{i=1}^{d} \boldsymbol{\omega}_{0,i}(\sigma_i^2 + \mu_i^2)\right] + \Psi, \tag{74}$$

where $\Psi = -\frac{1}{2}\left[d\log\phi(\mathrm{SSDI}) + \sum_{i=1}^{d}\log(\gamma\boldsymbol{\omega}_{0,i}\sigma_i^2) + d\right]$ captures the model complexity.

Decomposing the summation into label-related and client-related components

$$\sum_{i=1}^{d} \boldsymbol{\omega}_{0,i}(\sigma_i^2 + \mu_i^2) = \sum_{c=1}^{C} \frac{N_{\mathrm{total}}}{N_c}(\sigma_c^2 + \mu_c^2) + \sum_{k=1}^{K} \frac{N_{\mathrm{total}}}{n_k}(\sigma_k^2 + \mu_k^2). \tag{75}$$

This gives the final penalty terms

$$\Gamma_{\mathrm{LDS}} = \frac{\gamma}{2}\phi(\mathrm{SSDI}) \sum_{c=1}^{C} \frac{N_{\mathrm{total}}}{N_c}(\sigma_c^2 + \mu_c^2), \tag{76}$$

$$\Gamma_{\mathrm{LCD}} = \frac{\gamma}{2}\phi(\mathrm{SSDI}) \sum_{k=1}^{K} \frac{N_{\mathrm{total}}}{n_k}(\sigma_k^2 + \mu_k^2). \tag{77}$$

Substituting into the PAC-Bayes inequality completes the proof of Theorem 1. Here, μ_c^2, μ_k^2, $\sigma_{q,c}^2$, and $\sigma_{q,k}^2$ denote effective grouped terms rather than individual parameter coordinates. They summarize the posterior drift and uncertainty associated with the class-weighted and client-weighted parts of ω_0, and are used to express the KL penalty in a structured heterogeneity-aware form. □

B.5.2 Lemma 2: KL divergence under adaptive prior

Lemma 2. *For the adaptive prior* $P(\boldsymbol{\theta}) = \mathcal{N}(0, \boldsymbol{\Sigma}_p)$ *with* $\boldsymbol{\Sigma}_p^{-1} = \gamma \cdot \mathrm{Diag}(\phi(SSDI) \cdot \boldsymbol{\omega}_0)$ *and posterior* $Q(\boldsymbol{\theta}) = \mathcal{N}(\mu, \boldsymbol{\Sigma}_q)$*, the KL divergence is*

$$D_{KL}(Q\|P) = \frac{1}{2}\left[\gamma\phi(SSDI)\sum_{i=1}^{d}\boldsymbol{\omega}_{0,i}(\sigma_i^2+\mu_i^2) - d\log\phi(SSDI) - \sum_{i=1}^{d}\log(\gamma\boldsymbol{\omega}_{0,i}\sigma_i^2) - d\right]. \tag{78}$$

Proof. For two multivariate Gaussian distributions $Q = \mathcal{N}(\mu, \boldsymbol{\Sigma}_q)$ and $P = \mathcal{N}(0, \boldsymbol{\Sigma}_p)$, the KL divergence is

$$D_{\mathrm{KL}}(Q\|P) = \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) + \mu^\top\boldsymbol{\Sigma}_p^{-1}\mu - \log\det(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) - d\right]. \tag{79}$$

Given the diagonal structure $\boldsymbol{\Sigma}_p^{-1} = \gamma \cdot \mathrm{Diag}(\phi(\mathrm{SSDI}) \cdot \boldsymbol{\omega}_0)$ and $\boldsymbol{\Sigma}_q = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$, we compute each term

$$\mathrm{tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) = \gamma\phi(\mathrm{SSDI})\sum_{i=1}^{d}\boldsymbol{\omega}_{0,i}\sigma_i^2, \tag{80}$$

$$\mu^\top\boldsymbol{\Sigma}_p^{-1}\mu = \gamma\phi(\mathrm{SSDI})\sum_{i=1}^{d}\boldsymbol{\omega}_{0,i}\mu_i^2, \tag{81}$$

$$\log\det(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q) = \sum_{i=1}^{d}\log(\gamma\phi(\mathrm{SSDI})\boldsymbol{\omega}_{0,i}\sigma_i^2). \tag{82}$$

Substituting these terms

$$D_{\mathrm{KL}}(Q\|P) = \frac{1}{2}\left[\gamma\phi(\mathrm{SSDI})\sum_{i=1}^{d}\boldsymbol{\omega}_{0,i}(\sigma_i^2+\mu_i^2) - \sum_{i=1}^{d}\log(\gamma\phi(\mathrm{SSDI})\boldsymbol{\omega}_{0,i}\sigma_i^2) - d\right]. \tag{83}$$

Separating the $\log\phi(\mathrm{SSDI})$ term

$$\sum_{i=1}^{d}\log(\gamma\phi(\mathrm{SSDI})\boldsymbol{\omega}_{0,i}\sigma_i^2) = d\log\phi(\mathrm{SSDI}) + \sum_{i=1}^{d}\log(\gamma\boldsymbol{\omega}_{0,i}\sigma_i^2). \tag{84}$$

Thus, the final expression is

$$D_{\mathrm{KL}}(Q\|P) = \frac{1}{2}\left[\gamma\phi(\mathrm{SSDI})\sum_{i=1}^{d}\boldsymbol{\omega}_{0,i}(\sigma_i^2+\mu_i^2) - d\log\phi(\mathrm{SSDI}) - \sum_{i=1}^{d}\log(\gamma\boldsymbol{\omega}_{0,i}\sigma_i^2) - d\right]. \tag{85}$$

□

B.5.3 Derivation of Penalty Terms

The penalty terms derived in Theorem 1 capture two distinct heterogeneity effects:

Γ_{LDS}: Penalty due to Label Distribution Skew, quantifying the risk from non-uniform distribution of present labels across clients.

Γ_{LCD}: Penalty due to Label Coverage Deficiency, quantifying the risk from systematic label missingness.

Under the model assumptions from Section 4.1 of the main paper:

Parameter drift: $\mu_k^2 \propto \sum_{c=1}^{C} M_{c,k}N_c$.

Posterior variance: $\sigma_k^2 \propto n_k^{-1}$.

Global label distribution: $N_c \sim \text{Zipf}(\beta = 1)$.

Client data sizes: $n_k \sim \text{Pareto}(\alpha = 2)$.

The physical interpretation of these penalty terms aligns with the mechanisms described in Proposition 1 of the main paper: Label Distribution Skew creates a gradient of performance across labels, while Label Coverage Deficiency triggers polarized performance patterns with near-zero accuracy for missing labels.

The detailed asymptotic analysis of these penalty terms and the derivation of the critical scaling condition are provided in Appendix B.6 and Appendix B.7.

B.6 Proof of Theorem 2 and corollaries

B.6.1 Preliminaries and Assumptions

The asymptotic analysis builds upon the core model assumptions stated in Section 4.1 of the main text. To facilitate tractable analysis of the scaling laws governing the heterogeneity penalty terms, we introduce one additional statistical assumption.

Assumption 5. *Client data sizes n_k and label missingness indicators $\{M_{c,k}\}$ are statistically independent. That is, $n_k \perp \{M_{c,k}\}$ for all clients k and classes c.*

This assumption serves to decouple the effects of client-scale heterogeneity from label-coverage heterogeneity in our analysis. It enables a clean separation of the expectations of terms involving $n_k^{-1} M_{c,k}$, which is crucial for deriving the precise asymptotic orders.

B.6.2 Proof of Theorem 2

Proof. We prove the asymptotic orders for both Γ_{LCD} and Γ_{LDS} under the stated assumptions.

Part 1: Asymptotic Order of Γ_{LCD}

Recall the LCD penalty term

$$\Gamma_{\text{LCD}} = \frac{\gamma}{2} \phi(\text{SSDI}) \sum_{k=1}^{K} \frac{N_{\text{total}}}{n_k} (\sigma_k^2 + \mu_k^2). \tag{86}$$

Under the model assumptions

$$\sigma_k^2 \propto n_k^{-1}, \quad \mu_k^2 \propto \sum_{c=1}^{C} M_{c,k} N_c, \quad M_{c,k} \perp n_k, \tag{87}$$

the LCD penalty decomposes into two terms

$$\Gamma_{\text{LCD}} = (1 - \alpha) \left[N_{\text{total}} \sum_{k=1}^{K} n_k^{-2} + N_{\text{total}} \sum_{k=1}^{K} n_k^{-1} \sum_{c=1}^{C} M_{c,k} N_c \right], \tag{88}$$

For Term A ($N_{\text{total}} \sum_{k=1}^{K} n_k^{-2}$), under $n_k \sim \text{Pareto}(x_m, \alpha = 2)$ with $x_m = \frac{N_{\text{total}}}{2K}$,

$$\mathbb{E}[n_k^{-2}] = \frac{2K^2}{N_{\text{total}}^2}, \quad \mathbb{E}\left[\sum_{k=1}^{K} n_k^{-2} \right] = \frac{2K^3}{N_{\text{total}}^2}, \tag{89}$$

$$\mathbb{E}[\text{Term A}] = N_{\text{total}} \cdot \frac{2K^3}{N_{\text{total}}^2} = \frac{2K^3}{N_{\text{total}}}. \tag{90}$$

For Term B ($N_{\text{total}} \sum_{k=1}^{K} n_k^{-1} \sum_{c=1}^{C} M_{c,k} N_c$), under independence

$$\mathbb{E}[n_k^{-1}] = \frac{4K}{3N_{\text{total}}}, \quad \mathbb{E}[\text{Term B}] = \frac{4K^2}{3} \sum_{c=1}^{C} p_c = O(CK^2), \tag{91}$$

combine both terms

$$\Gamma_{\text{LCD}} = (1-\alpha)\left(\lambda_1 \cdot \frac{K^3}{N_{\text{total}}} + \lambda_2 \cdot CK^2\right) \sim O(CK^2). \tag{92}$$

Part 2: Asymptotic Order of Γ_{LDS}

Recall the LDS penalty term

$$\Gamma_{\text{LDS}} = \frac{\gamma}{2}\phi(\text{SSDI}) \sum_{c=1}^{C} \frac{N_{\text{total}}}{N_c}(\sigma_c^2 + \mu_c^2). \tag{93}$$

Under Zipf distribution $N_c \sim \text{Zipf}(\beta = 1)$ with normalisation

$$N_c = A \cdot c^{-1}, \quad A = \frac{N_{\text{total}}}{H_C} \approx \frac{N_{\text{total}}}{\ln C}. \tag{94}$$

The LDS penalty decomposes into two terms

$$\Gamma_{\text{LDS}} = \alpha \left[N_{\text{total}} \sum_{c=1}^{C} N_c^{-2} + N_{\text{total}} \sum_{c=1}^{C} \frac{1}{N_c \eta_c} \right]. \tag{95}$$

For Term C ($N_{\text{total}} \sum_{c=1}^{C} N_c^{-2}$),

$$\sum_{c=1}^{C} N_c^{-2} = A^{-2} \sum_{c=1}^{C} c^2 \approx A^{-2} \cdot \frac{C^3}{3} = \frac{(\ln C)^2 C^3}{3N_{\text{total}}^2}, \tag{96}$$

$$\text{Term C} = N_{\text{total}} \cdot \frac{(\ln C)^2 C^3}{3N_{\text{total}}^2} = \frac{(\ln C)^2 C^3}{3N_{\text{total}}}. \tag{97}$$

For Term D ($N_{\text{total}} \sum_{c=1}^{C} \frac{1}{N_c \eta_c}$), using coverage assumption $\eta_c = \kappa(N_c/N_{\text{total}})$,

$$\frac{1}{N_c \eta_c} = \frac{1}{\kappa} \cdot \frac{N_{\text{total}}}{N_c^2}, \tag{98}$$

$$\text{Term D} = \frac{(\ln C)^2 C^3}{3\kappa}. \tag{99}$$

Combining both terms

$$\Gamma_{\text{LDS}} = \alpha \left(\lambda_3 \cdot \frac{(\ln C)^2 C^3}{N_{\text{total}}} + \lambda_4 \cdot (\ln C)^2 C^3 \right) \sim O(C^3 \ln^2 C). \tag{100}$$

This completes the proof of Theorem 2. □

B.6.3 Proof of Corollary 2

Proof. The critical scaling condition is derived by comparing the dominant terms of Γ_{LCD} and Γ_{LDS}.

From Theorem 3 and Theorem 4, the dominant terms satisfy:

$$\Gamma_{\text{LCD}} \sim O(CK^2), \quad \Gamma_{\text{LDS}} \sim O(C^3 \ln^2 C). \tag{101}$$

Now we incorporate the system-specific constants from the detailed derivations. Using the detailed derivation in Appendix B.7.1, the LCD penalty can be written as

$$\Gamma_{\text{LCD}} = (1-\alpha)\left(\lambda_1 \cdot \frac{K^3}{N_{\text{total}}} + \lambda_2 \cdot CK^2\right),$$

where $\lambda_2 = \frac{4}{3}\bar{p}$ and $\bar{p} = \frac{1}{C}\sum_{c=1}^{C} p_c$ is the average missing probability.

Using the detailed derivation in Appendix B.7.2, the LDS penalty can be written as

$$\Gamma_{\text{LDS}} = \alpha\left(\lambda_3 \cdot \frac{(\ln C)^2 C^3}{N_{\text{total}}} + \lambda_4 \cdot (\ln C)^2 C^3\right),$$

where $\lambda_4 = \frac{1}{3\kappa_c}$ and κ_c is the coverage constant from the assumption $\eta_c = \kappa_c(N_c/N_{\text{total}})$.

Comparing the dominant terms $\lambda_2 CK^2$ and $\lambda_4(\ln C)^2 C^3$,

$$\begin{aligned}
\lambda_2 CK^2 &\sim \lambda_4 (\ln C)^2 C^3, \\
\frac{4}{3}\bar{p} CK^2 &\sim \frac{1}{3\kappa_c}(\ln C)^2 C^3, \\
K^2 &\sim \frac{1}{4\kappa_c \bar{p}} C^2 (\ln C)^2.
\end{aligned}$$

Taking square roots and incorporating the balance parameter α,

$$K_{\text{crit}} = \sqrt{\frac{\alpha}{4(1-\alpha)\kappa_c \bar{p}}} \cdot C \ln C.$$

Let $\kappa = \sqrt{\frac{\alpha}{4(1-\alpha)\kappa_c \bar{p}}}$ be the system constant that aggregates all system-specific factors, then

$$K_{\text{crit}} = \kappa \cdot C \ln C.$$

Parameter Interpretation:

K_{crit}: Critical number of clients where dominance transitions from LDS to LCD.

α: Balance parameter between LDS and LCD components ($0 < \alpha < 1$).

κ_c: Coverage constant from $\eta_c = \kappa_c(N_c/N_{\text{total}})$, representing label coverage efficiency.

$\bar{p}$: Average missing probability, $\bar{p} = \frac{1}{C}\sum_{c=1}^{C} p_c$.

κ: System-specific constant aggregating α, κ_c, and $\bar{p}$.

The three distinct phases are:

LDS-Dominant Phase ($K \ll K_{\text{crit}}$): Characterised by imbalanced learning but all labels learnable. In this regime, Label Distribution Skew creates performance gradients but maintains learning opportunities for all classes.

Transition Phase ($K \approx K_{\text{crit}}$): Mixed behavior requiring balanced strategies. Both heterogeneity types contribute comparably to the overall risk, necessitating comprehensive mitigation approaches.

LCD-Dominant Phase ($K \gg K_{\text{crit}}$): Characterized by catastrophic failure on missing labels. Label Coverage Deficiency becomes the dominant risk factor, leading to near-zero accuracy for systematically missing classes.

This critical scaling analysis provides fundamental insights into federated learning system design: client fragmentation primarily amplifies label coverage issues, while label diversity exacerbates distribution skew. The transition point $K_{\text{crit}} \propto C \ln C$ offers concrete guidance for system capacity planning and heterogeneity management. □

B.7 Detailed Derivations for Asymptotic Analysis

B.7.1 Detailed derivation for the asymptotic order of Γ_{LCD}

We provide the detailed asymptotic derivation under the stated assumptions for the asymptotic order of Γ_{LCD}.

Theorem 3. *Under the assumptions of Pareto-distributed client data sizes and independence between label missingness and data sizes, the Label Coverage Deficiency penalty has asymptotic order $\Gamma_{LCD} \sim O(CK^2)$ with detailed decomposition:*

$$\Gamma_{LCD} = (1-\alpha)\left[N_{total}\sum_{k=1}^{K} n_k^{-2} + N_{total}\sum_{k=1}^{K} n_k^{-1}\sum_{c=1}^{C} M_{c,k}N_c\right]. \tag{102}$$

Proof. **Part 1: Term A Derivation**

Under $n_k \sim \text{Pareto}(x_m, \alpha = 2)$ with $x_m = \frac{N_{\text{total}}}{2K}$, the probability density function is

$$f_{n_k}(x) = \frac{2x_m^2}{x^3}, \quad \text{for} \quad x \geq x_m. \tag{103}$$

Compute the expectation

$$\mathbb{E}[n_k^{-2}] = \int_{x_m}^{\infty} x^{-2} \cdot \frac{2x_m^2}{x^3} dx = 2x_m^2 \int_{x_m}^{\infty} x^{-5} dx = 2x_m^2 \left[-\frac{1}{4}x^{-4}\right]_{x_m}^{\infty} = \frac{1}{2}x_m^{-2}. \tag{104}$$

Substituting $x_m = \frac{N_{\text{total}}}{2K}$,

$$\mathbb{E}[n_k^{-2}] = \frac{1}{2}\left(\frac{2K}{N_{\text{total}}}\right)^2 = \frac{2K^2}{N_{\text{total}}^2}. \tag{105}$$

For K independent clients,

$$\mathbb{E}\left[\sum_{k=1}^{K} n_k^{-2}\right] = K \cdot \frac{2K^2}{N_{\text{total}}^2} = \frac{2K^3}{N_{\text{total}}^2}. \tag{106}$$

Thus,

$$\mathbb{E}[\text{Term A}] = N_{\text{total}} \cdot \frac{2K^3}{N_{\text{total}}^2} = \frac{2K^3}{N_{\text{total}}}. \tag{107}$$

Part 2: Term B Derivation

Under independence $n_k \perp \{M_{c,k}\}$, first compute $\mathbb{E}[n_k^{-1}]$,

$$\mathbb{E}[n_k^{-1}] = \int_{x_m}^{\infty} x^{-1} \cdot \frac{2x_m^2}{x^3} dx = 2x_m^2 \int_{x_m}^{\infty} x^{-4} dx = 2x_m^2 \left[-\frac{1}{3}x^{-3}\right]_{x_m}^{\infty} = \frac{2}{3}x_m^{-1}. \tag{108}$$

Substituting $x_m = \frac{N_{\text{total}}}{2K}$,

$$\mathbb{E}[n_k^{-1}] = \frac{2}{3} \cdot \frac{2K}{N_{\text{total}}} = \frac{4K}{3N_{\text{total}}}. \tag{109}$$

Now compute Term B expectation

$$\mathbb{E}[\text{Term B}] = \mathbb{E}\left[N_{\text{total}} \sum_{k=1}^{K} n_k^{-1} \sum_{c=1}^{C} M_{c,k} N_c\right] = N_{\text{total}} \sum_{k=1}^{K} \sum_{c=1}^{C} \mathbb{E}[M_{c,k}] \cdot \mathbb{E}[n_k^{-1}]. \tag{110}$$

Let $p_c = \mathbb{E}[M_{c,k}]$ be the missing probability for class c, then $\sum_{c=1}^{C} M_{c,k} N_c \approx \frac{N_{\text{total}}}{C} \sum_{c=1}^{C} M_{c,k}$ under the Zipf-based average class-scale approximation used for order analysis,

$$\mathbb{E}[\text{Term B}] = N_{\text{total}} \sum_{k=1}^{K} \sum_{c=1}^{C} p_c \cdot \frac{4K}{3N_{\text{total}}} = \frac{4K}{3} \sum_{c=1}^{C} p_c \sum_{k=1}^{K} 1 = \frac{4K^2}{3} \sum_{c=1}^{C} p_c. \tag{111}$$

Assuming $\sum_{c=1}^{C} p_c = O(C)$, we obtain

$$\mathbb{E}[\text{Term B}] = O(CK^2). \tag{112}$$

Part 3: Combined LCD Penalty

Combining both terms with the balance parameter $(1-\alpha)$,

$$\Gamma_{\text{LCD}} = (1-\alpha)(\text{Term A} + \text{Term B}) = (1-\alpha)\left(\lambda_1 \cdot \frac{K^3}{N_{\text{total}}} + \lambda_2 \cdot CK^2\right). \tag{113}$$

For large K and realistic federated learning conditions where $N_{\text{total}} \gg K \cdot C$, Term B dominates

$$\Gamma_{\text{LCD}} \sim O(CK^2). \tag{114}$$

□

B.7.2 Detailed derivation for the asymptotic order of Γ_{LDS}

We provide the detailed asymptotic derivation under the stated assumptions for the asymptotic order of Γ_{LDS}.

Theorem 4. *Under the assumptions of Zipf-distributed global label distribution and label coverage proportional to global sample size, the Label Distribution Skew penalty has asymptotic order $\Gamma_{LDS} \sim O(C^3 \ln^2 C)$ with detailed decomposition:*

$$\Gamma_{LDS} = \alpha \left[N_{total} \sum_{c=1}^{C} N_c^{-2} + N_{total} \sum_{c=1}^{C} \frac{1}{N_c \eta_c}\right]. \tag{115}$$

Proof. **Part 1: Zipf Distribution Preliminaries**

Under $N_c \sim \text{Zipf}(\beta = 1)$ with normalisation,

$$N_c = A \cdot c^{-1}, \quad A = \frac{N_{\text{total}}}{H_C} \approx \frac{N_{\text{total}}}{\ln C}, \tag{116}$$

where $H_C = \sum_{c=1}^{C} c^{-1} \approx \ln C + \gamma$ is the harmonic number.

Part 2: Term C Derivation

Compute the sum of inverse squared class sizes

$$\sum_{c=1}^{C} N_c^{-2} = \sum_{c=1}^{C} (A \cdot c^{-1})^{-2} = A^{-2} \sum_{c=1}^{C} c^2. \tag{117}$$

Using the formula for the sum of squares

$$\sum_{c=1}^{C} c^2 = \frac{C(C+1)(2C+1)}{6} \approx \frac{C^3}{3}, \quad \text{for large } C. \tag{118}$$

Thus,

$$\sum_{c=1}^{C} N_c^{-2} \approx A^{-2} \cdot \frac{C^3}{3}. \tag{119}$$

Substituting $A \approx \frac{N_{\text{total}}}{\ln C}$,

$$\sum_{c=1}^{C} N_c^{-2} \approx \left(\frac{\ln C}{N_{\text{total}}}\right)^2 \cdot \frac{C^3}{3} = \frac{(\ln C)^2 C^3}{3N_{\text{total}}^2}. \tag{120}$$

Therefore,

$$\text{Term C} = N_{\text{total}} \sum_{c=1}^{C} N_c^{-2} \approx N_{\text{total}} \cdot \frac{(\ln C)^2 C^3}{3N_{\text{total}}^2} = \frac{(\ln C)^2 C^3}{3N_{\text{total}}}. \tag{121}$$

Part 3: Term D Derivation

Using the coverage assumption $\eta_c = \kappa(N_c/N_{\text{total}})$,

$$\frac{1}{N_c \eta_c} = \frac{1}{N_c \cdot (\kappa N_c / N_{\text{total}})} = \frac{1}{\kappa} \cdot \frac{N_{\text{total}}}{N_c^2}. \tag{122}$$

Now compute

$$\text{Term D} = N_{\text{total}} \sum_{c=1}^{C} \frac{1}{N_c \eta_c} = N_{\text{total}} \sum_{c=1}^{C} \frac{1}{\kappa} \cdot \frac{N_{\text{total}}}{N_c^2} = \frac{N_{\text{total}}^2}{\kappa} \sum_{c=1}^{C} N_c^{-2}. \tag{123}$$

Using the previous result for $\sum_{c=1}^{C} N_c^{-2}$,

$$\text{Term D} = \frac{N_{\text{total}}^2}{\kappa} \cdot \frac{(\ln C)^2 C^3}{3N_{\text{total}}^2} = \frac{(\ln C)^2 C^3}{3\kappa}. \tag{124}$$

Part 4: Combined LDS Penalty

Combining both terms with the balance parameter α

$$\Gamma_{\text{LDS}} = \alpha(\text{Term C} + \text{Term D}) = \alpha \left(\lambda_3 \cdot \frac{(\ln C)^2 C^3}{N_{\text{total}}} + \lambda_4 \cdot (\ln C)^2 C^3\right). \tag{125}$$

For typical federated learning scenarios, Term D dominates

$$\Gamma_{\text{LDS}} \sim O(C^3 \ln^2 C). \tag{126}$$

□

B.7.3 Dominance Hierarchy Analysis

The dominance hierarchy reveals key insights for federated learning system design:

1. **Client fragmentation** (K increase) primarily amplifies label coverage issues through the LCD penalty.
2. **Label diversity** (C increase) primarily exacerbates distribution skew through the LDS penalty.
3. Data volume scaling has a limited effect on structural heterogeneity, as both penalties show weak dependence on N_{total}.
4. The critical scaling $K \sim C \ln C$ provides concrete guidance for system capacity planning.

The dominance hierarchy remains robust under moderate variations in distributional parameters, and the critical scaling relationship provides a conservative baseline for system design.

B.8 Theoretical Motivation via OU Process

B.8.1 Motivation and OU Process Framework

In this section, we employ the Ornstein-Uhlenbeck (OU) process framework to provide theoretical justification for Assumptions 3 and 4 stated in the main text. While this does not constitute a fully rigorous proof in the most general setting, this analysis demonstrates that these assumptions emerge naturally from fundamental stochastic process modelling of federated learning dynamics.

The OU process captures the essential trade-off between client-specific stochastic drift (due to local data heterogeneity) and global model aggregation (which pulls parameters toward a common value).

B.8.2 From Discrete to Continuous Dynamics

We begin by modelling the evolution of client parameters in federated learning. Consider the local stochastic gradient descent (SGD) update for client k at communication round t,

$$\boldsymbol{\theta}_k^{(t+1)} = \boldsymbol{\theta}_k^{(t)} - \eta \left[\nabla \mathcal{L}_k(\boldsymbol{\theta}_k^{(t)}) + \boldsymbol{\xi}_k^{(t)} \right], \tag{127}$$

where $\boldsymbol{\xi}_k^{(t)}$ is the stochastic gradient noise with $\mathbb{E}[\boldsymbol{\xi}_k^{(t)}] = 0$ and $\mathrm{Cov}(\boldsymbol{\xi}_k^{(t)}) = \boldsymbol{\Sigma}_k$.

Proposition 4 (Gradient Noise Scaling)**.** *Under standard assumptions of bounded gradients and IID sampling within each client, the covariance of stochastic gradient noise scales inversely with local data size:*

$$Tr(\boldsymbol{\Sigma}_k) = O\left(\frac{1}{n_k}\right). \tag{128}$$

Proof. For empirical risk minimisation with n_k IID samples, the central limit theorem implies that gradient estimation error variance decreases as $O(1/n_k)$.

Specifically, for a loss function $\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{n_k}\sum_{i=1}^{n_k} \ell(\boldsymbol{\theta}; z_i)$, we have

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k^2} \sum_{i=1}^{n_k} \mathrm{Cov}(\nabla \ell(\boldsymbol{\theta}; z_i)) = O\left(\frac{1}{n_k}\right) I, \tag{129}$$

where individual gradient covariances are bounded by standard smoothness assumptions. □

Taking the continuous-time limit with learning rate $\eta \to 0$ and number of local steps τ, we obtain the Ornstein-Uhlenbeck process

$$d\boldsymbol{\theta}_k(t) = -\lambda(\boldsymbol{\theta}_k(t) - \bar{\boldsymbol{\theta}})dt + \sqrt{\eta\tau\boldsymbol{\Sigma}_k}dW_t, \tag{130}$$

where λ represents the mean-reversion strength due to periodic model aggregation, and W_t is a standard Wiener process.

B.8.3 Proposition 4: justification for Assumption 4

Proposition 5. *The OU process framework provides theoretical justification for Assumption 4: posterior variance scales inversely with local data size, $\sigma_{q,k}^2 \propto n_k^{-1}$.*

Proof. Consider the OU process

$$d\boldsymbol{\theta}_k(t) = -\lambda(\boldsymbol{\theta}_k(t) - \bar{\boldsymbol{\theta}})dt + \sigma_k dW_t. \tag{131}$$

This process has a stationary distribution given by

$$\boldsymbol{\theta}_k(\infty) \sim \mathcal{N}\left(\bar{\boldsymbol{\theta}}, \frac{\sigma_k^2}{2\lambda} I\right). \tag{132}$$

From Proposition 2, we have $\boldsymbol{\Sigma}_k \propto \frac{1}{n_k} I$. In the continuous-time limit, the effective diffusion coefficient scales as

$$\sigma_k^2 = \eta\tau \mathrm{Tr}(\boldsymbol{\Sigma}_k) \propto \frac{1}{n_k}. \tag{133}$$

Apply the stationary distribution result,

$$\sigma_{q,k}^2 = \mathrm{Var}[\boldsymbol{\theta}_k(\infty)] = \frac{\sigma_k^2}{2\lambda} \propto \frac{1}{n_k}. \tag{134}$$

This provides strong theoretical justification for Assumption 4, showing that the inverse relationship between posterior variance and local data size emerges directly from the fundamental scaling properties of stochastic gradient estimation. □

B.8.4 Justification and surrogate rationale for Assumption 3

The OU process framework provides qualitative support for the principle underlying Assumption 3: client-level drift increases with label missingness, especially when globally more prevalent labels are absent.

We first define the gradient bias from missing labels. For client k missing label c, the expected gradient bias is

$$\mathbf{b}_{k,c} = \mathbb{E}_{z\sim P_c}[\nabla \ell(\boldsymbol{\theta}; z)] - \mathbb{E}_{z\sim P_k}[\nabla \ell(\boldsymbol{\theta}; z)], \tag{135}$$

where P_c is the data distribution for class c and P_k is client k's data distribution.

The total gradient error decomposes as

$$\nabla \mathcal{L}_k(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta}) = \underbrace{\sum_{c=1}^{C} M_{c,k} \rho_c^{\text{global}} \mathbf{b}_{k,c}}_{\text{systematic bias}} + \underbrace{\boldsymbol{\xi}_k^{(t)}}_{\text{stochastic noise}}. \tag{136}$$

In the presence of missing labels, the parameter evolution follows the modified OU process

$$d\boldsymbol{\theta}_k(t) = -\lambda(\boldsymbol{\theta}_k(t) - \bar{\boldsymbol{\theta}})dt + \mathbf{B}_k dt + \sqrt{\eta\tau}\boldsymbol{\Sigma}_k dW_t, \tag{137}$$

where $\mathbf{B}_k = \sum_{c=1}^{C} M_{c,k}\rho_c^{\text{global}}\mathbf{b}_{k,c}$ is the aggregate bias from missing labels.

This modified OU process has a stationary distribution

$$\boldsymbol{\theta}_k(\infty) \sim \mathcal{N}\left(\bar{\boldsymbol{\theta}} + \frac{\mathbf{B}_k}{\lambda}, \frac{\eta\tau\boldsymbol{\Sigma}_k}{2\lambda} I\right). \tag{138}$$

The parameter drift in expectation is

$$\mu_k = \mathbb{E}[\boldsymbol{\theta}_k(\infty)] - \bar{\boldsymbol{\theta}} = \frac{\mathbf{B}_k}{\lambda} = \frac{1}{\lambda}\sum_{c=1}^{C} M_{c,k}\rho_c^{\text{global}}\mathbf{b}_{k,c}. \tag{139}$$

Taking squared norm and expectation

$$\mathbb{E}[\|\mu_k\|^2] = \frac{1}{\lambda^2}\mathbb{E}\left[\left\|\sum_{c=1}^{C} M_{c,k}\rho_c^{\text{global}}\mathbf{b}_{k,c}\right\|^2\right]. \tag{140}$$

Under the physically motivated approximations that:

1. The norm of gradient bias scales with missing class prevalence: $\|\mathbf{b}_{k,c}\| \propto \rho_c^{\text{global}} \propto N_c$.
2. Bias directions for different classes are approximately orthogonal: $\mathbf{b}_{k,c}^T\mathbf{b}_{k,c'} \approx 0$ for $c \neq c'$.

We obtain

$$\mathbb{E}[\|\mu_k\|^2] \approx \frac{1}{\lambda^2}\sum_{c=1}^{C} M_{c,k}(\rho_c^{\text{global}})^2\mathbb{E}[\|\mathbf{b}_{k,c}\|^2] \propto \sum_{c=1}^{C} M_{c,k}(N_c/N_{\text{total}})^2 \cdot N_c^2 = \sum_{c=1}^{C} M_{c,k}N_c^4/N_{\text{total}}^2. \tag{141}$$

This expression is not intended to yield Assumption 3 exactly. Rather, it suggests the key structural principle that client-level drift increases when labels with larger global prevalence are absent.

Motivated by this monotonic dependence, we adopt the prevalence-weighted score

$$\sum_{c=1}^{C} M_{c,k} N_c \tag{142}$$

as a surrogate descriptor of client-level drift. Accordingly, Assumption 3 should be interpreted as a modelling choice that preserves the prevalence-aware effect of missing labels, rather than as a strict closed-form consequence of the OU analysis.

B.8.5 Normalized Approximation and Limitations

For the asymptotic order analysis in Appendix B.7, it is useful to further replace the prevalence-weighted score by a coarse-grained normalized approximation in which class-specific prevalence is represented by the average class scale. Under this approximation,

$$\sum_{c=1}^{C} M_{c,k} N_c \approx \frac{N_{\text{total}}}{C} \sum_{c=1}^{C} M_{c,k}. \tag{143}$$

This approximation is not intended to preserve class-specific prevalence effects exactly. Rather, it provides an average-scale surrogate that is sufficient for the coarse-grained order analysis, and recovers an effective linear dependence on the number of missing classes at the client level.

Accordingly, the asymptotic derivation of Γ_{LCD} should be interpreted as operating on this normalized surrogate form, rather than on the unnormalized prevalence-weighted score directly.

We acknowledge several idealisations in the OU process derivations:

Continuous-Time Approximation: The discrete-time nature of federated learning is approximated by a continuous-time process.

Linear Dynamics: The OU process assumes linear dynamics, while deep learning optimisation is inherently non-convex and non-linear.

Gaussian Assumptions: Both the gradient noise and the resulting parameter distributions are assumed to be Gaussian.

Coarse-Grained Drift Approximation: The normalized approximation above is introduced for asymptotic tractability and does not preserve fine-grained class-specific prevalence effects exactly.

Simplified Bias Modelling: The specific form $\|\mathbf{b}_{k,c}\| \propto N_c$ and the approximate orthogonality assumption are modelling simplifications.

Despite these limitations, the OU process framework still provides useful qualitative support for the surrogate assumptions used in the heterogeneity penalty analysis.

B.8.6 Summary of the OU-Based Motivation

Overall, the Ornstein-Uhlenbeck process framework provides a useful interpretive basis for the two auxiliary assumptions used in the heterogeneity penalty analysis:

Assumption 4 is supported by the standard $O(1/n_k)$ scaling of SGD noise variance with local sample size.

Assumption 3 is supported at the qualitative level by the observation that missing more globally prevalent labels induces larger client-level drift; for asymptotic analysis, this effect is represented through a surrogate prevalence-weighted score and its normalized approximation.

Therefore, the OU analysis should be viewed primarily as a source of qualitative support and modelling rationale, rather than as a strict derivation of the heterogeneity penalty terms.

C Supplementary Experimental Results and Analyses

C.1 Experimental Setup Details

C.1.1 Dataset and Model Specifications

Dataset details:

MNIST: 70,000 grayscale handwritten digit images from 10 classes, split into 60,000 training and 10,000 test samples. Standard normalisation uses mean 0.1307 and standard deviation 0.3081.

CIFAR-100 (coarse): 60,000 colour images grouped into 20 coarse categories from the original 100 fine-grained classes. Data augmentation includes random cropping and horizontal flipping. Standard normalisation uses mean $[0.5071, 0.4867, 0.4408]$ and standard deviation $[0.2675, 0.2565, 0.2761]$.

Tiny ImageNet: 200,000 images from 200 classes, resized to 64×64. Standard normalisation uses mean $[0.5, 0.5, 0.5]$ and standard deviation $[0.5, 0.5, 0.5]$.

Table 4 summarises the basic specifications of the datasets used in our experiments, including dataset size, number of classes, and input dimension.

Table 4: Dataset specifications used in experiments

Dataset	Size	Classes	Input Dimension
MNIST	70,000	10	$1 \times 28 \times 28$
CIFAR-100 (coarse)	60,000	20	$3 \times 32 \times 32$
Tiny ImageNet (subset)	200,000	200	$3 \times 64 \times 64$

Model architectures:

Table 5 summarises the model architectures adopted for different datasets, including the primary backbone used in the main experiments and alternative models used in supplementary evaluations.

Table 5: Model architectures used for different datasets

Dataset	Primary Model	Alternative Model
MNIST	2-Layer Neural Network	-
CIFAR-100 (coarse)	ShuffleNetV2	ResNet-18
Tiny ImageNet	ResNet-18	-

2-Layer Neural Network (MNIST): Two fully connected layers with 200 hidden units each and ReLU activations, with input dimension 784 (flattened 28×28 images).

ShuffleNetV2: A lightweight architecture with channel-shuffle operations, used for efficient training on CIFAR-100 (coarse).

ResNet-18: A standard 18-layer residual network with GroupNorm replacing BatchNorm to improve stability under federated training.

C.1.2 Hyperparameters and Training Configuration

Table 6 summarises the main training hyperparameters used for the evaluated datasets, including optimisation settings, client participation ratio, and missing-rate configurations.

Training protocol:

Federated Averaging: We use the standard FedAvg algorithm with 10% client participation in each communication round.

Table 6: Training hyperparameters for the evaluated datasets.

Parameter	MNIST	CIFAR-100 (coarse)	Tiny ImageNet
Max Class	10	100(20)	200
Batch Size	128	128	128
Local Epochs	5	5	3
Optimizer	SGD	SGD	SGD
Learning Rate	0.01	0.05	0.05
Learning Rate Decay	0.002	0.005	0.01
Learning Rate Schedule	Exponential decay	Exponential decay	Exponential decay
Active Ratio	0.1	0.1	0.1
Missing Rate	0.30/0.5/0.7	0.30/0.5	0.30/0.5

Learning-rate scheduling: Exponential decay with factor 0.998 is applied after each communication round.

Client selection: Clients are sampled uniformly at random according to the client fraction, with fixed random seeds for reproducibility.

Local training: Each selected client performs local SGD updates with gradient clipping (maximum norm 10).

Aggregation: Client model parameters are averaged as

$$\boldsymbol{\theta}^{(t+1)} = \frac{1}{|\mathcal{S}_t|} \sum_{k \in \mathcal{S}_t} \boldsymbol{\theta}_k^{(t+1)}.$$

Evaluation Metrics:

Global Accuracy: Overall test accuracy across all classes.

Per-Class Accuracy: Individual accuracy for each class.

Client Divergence: Mean squared distance between client models and global model.

Divergence Statistics: Standard deviation, minimum, and maximum of client divergences per round.

C.1.3 Non-IID Data Partitioning Methodology

The complete controlled non-IID partition procedure used in our experiments is summarized in Algorithm 1.

The theoretical normalization in the main text allows zero-mass classes or zero-mass clients in the limiting extremal construction used to define $V_{\max}(C, K)$. By contrast, the actual data generator used in experiments enforces $N_c \geq 1$ and $n_k \geq 1$, so all classes and clients remain non-empty. As a result, when $C \neq K$, the empirical SSDI may not reach 1 even under extremely heterogeneous partitions, because the theoretical supremum is then not exactly attainable under the non-empty constraint.

Heterogeneity generation parameters:

Global label distribution: $N_c \sim \text{Zipf}(\alpha_z = 1.0)$, which yields a long-tailed label distribution that mimics real-world imbalance.

Client data-size distribution: $n_k \sim \text{Pareto}(\alpha_p = 2.0)$, which yields a heavy-tailed client-size distribution reflecting heterogeneous client capacities.

Missing-rate control: Fixed missing rates $MR \in \{0.3, 0.5, 0.7\}$ are used to control systematic label deficiencies.

Rare-label coverage: We explicitly impose lower coverage for rare labels to reflect realistic data scarcity patterns.

Algorithm 1 Non-IID Data Partitioning with Controlled Heterogeneity

Input: N_{total}, K, C, missing rate MR, Zipf parameter α_z, Pareto parameter α_p
Output: Client-class count matrix $\mathbf{N} = [n_{k,c}] \in \mathbb{Z}_{\geq 0}^{K \times C}$

Step 1. Global label distribution
$\{N_c\}_{c=1}^{C} \leftarrow \text{ZipfDistribution}(C, N_{\text{total}}, \alpha_z)$ ▷ sample class marginals
Adjust $\{N_c\}$ so that $\sum_{c=1}^{C} N_c = N_{\text{total}}$ and $N_c \geq 1$ ▷ enforce non-empty classes

Step 2. Client data sizes
$\{n_k\}_{k=1}^{K} \leftarrow \text{ParetoDistribution}(K, N_{\text{total}}, \alpha_p)$ ▷ sample client marginals
Adjust $\{n_k\}$ so that $\sum_{k=1}^{K} n_k = N_{\text{total}}$ and $n_k \geq 1$ ▷ enforce non-empty clients

Step 3. Missingness pattern
$\mathbf{M} = [M_{k,c}] \leftarrow \text{RandomMissingPattern}(K, C, MR)$ ▷ $M_{k,c} = 1$: class c missing on client k
Refine $\mathbf{M}$ to satisfy class/client feasibility and lower coverage for rare labels ▷ controlled LCD

Step 4. Sample allocation
Initialize $n_{k,c} \leftarrow 0$ for all (k, c)
$\Omega \leftarrow \{(k, c) : M_{k,c} = 0\}$ ▷ feasible pairs
while $\exists\, c : \sum_k n_{k,c} < N_c$ **or** $\exists\, k : \sum_c n_{k,c} < n_k$ **do**
  $\Omega_{\text{cur}} \leftarrow \{(k, c) \in \Omega : \sum_{k'} n_{k',c} < N_c,\ \sum_{c'} n_{k,c'} < n_k\}$ ▷ remaining feasible pairs
  $w_{k,c} \propto \text{ZipfWeight}(c)\, \text{ParetoWeight}(k),\ \forall (k, c) \in \Omega_{\text{cur}}$ ▷ allocation weights
  Sample $(k^*, c^*) \sim \Omega_{\text{cur}}$ according to $\{w_{k,c}\}$ ▷ single-step allocation
  $n_{k^*,c^*} \leftarrow n_{k^*,c^*} + 1$
  if stagnation is detected **then**
    Force allocation on residual feasible pairs; minimally relax $\mathbf{M}$ if needed ▷ resolve infeasibility
  end if
end while

Step 5. Validation
Verify $\sum_{k,c} n_{k,c} = N_{\text{total}}$, $\sum_k n_{k,c} = N_c$, and $\sum_c n_{k,c} = n_k$ ▷ marginal consistency
Compute $\epsilon_{\text{zipf}}, \epsilon_{\text{pareto}}, \epsilon_{MR}$; regenerate if any exceeds ϵ_0 ▷ validation tolerance
return $\mathbf{N} = [n_{k,c}]$

Validation of Partitioning:

Statistical Consistency: Verify $\sum_{k=1}^{K} n_k = \sum_{c=1}^{C} N_c = N_{\text{total}}$ and actual missing rate $\approx$ target MR.

Connectivity Guarantee: Ensure no empty clients or completely missing classes through constraint enforcement.

Distribution Fidelity: Monitor Zipf and Pareto parameter adherence throughout iterative optimisation.

Reproducibility: Fixed random seeds ensure consistent partitioning across experimental runs.

Methodology Advantages: The proposed partitioning creates realistic federated learning scenarios with:

Systemic Label Imbalance: Zipf-distributed global labels mimic long-tail phenomena in real applications.

Client Capacity Heterogeneity: Pareto-distributed client sizes reflect varying device capabilities and data collection opportunities.

Structured Missingness: Controlled missing rates with connectivity guarantees prevent training failures while maintaining realistic data deficiencies.

Configurable Heterogeneity: Adjustable parameters enable systematic study of federated learning under varying non-IID conditions.

C.2 SSDI Behavior Analysis

C.2.1 LDS vs LCD Components Across Client Counts

Experimental results:

Figure 1 illustrates SSDI and its components (LDS, LCD) across different client counts for 10 classes. Our analysis reveals systematic patterns in SSDI component behavior as client count increases:

Rapid initial decrease: Both overall SSDI and its LDS/LCD components decrease rapidly as K increases from small values ($K = 10$) to moderate values ($K = 50$–100).

Gradual plateau: After this initial decrease, the SSDI components approach a plateau, indicating diminishing sensitivity to further increases in the client count.

Consistency across missing rates: The same qualitative pattern is observed for missing rates 0.3, 0.5, and 0.7.

Interpretation:

Initial Decrease: The rapid initial decrease occurs because adding more clients initially helps distribute data more evenly, reducing both distribution skew and coverage deficiency.

Plateau Effect: The plateau emerges because beyond a certain point, further client fragmentation primarily exacerbates label coverage issues rather than improving distribution balance.

C.2.2 Deficiency-to-Skew Ratio Trends

DSR Definition and Significance: The Deficiency-to-Skew Ratio (DSR) is defined as

$$\mathrm{DSR} = \frac{\mathrm{SSDI}_{\mathrm{LCD}}}{\mathrm{SSDI}_{\mathrm{LDS}}}.$$

This ratio quantifies the relative dominance between Label Coverage Deficiency and Label Distribution Skew, providing crucial insights into the nature of heterogeneity in a given federated learning system.

U-Shaped Trend Analysis: Our experiments reveal a characteristic U-shaped trend in DSR:

Initial Decrease: DSR initially decreases as K increases from small values, indicating that distribution skew contributes a larger impact.

Minimum Point: The DSR reaches a minimum at K_{crit}, due to the synergistic interaction between the LDS and LCD effects.

Subsequent Increase: Beyond K_{crit}, DSR increases monotonically, indicating that the impact of missing labels grows more severe.

Critical Point Analysis:

Table 7 reports the empirical critical client counts K_{crit} under different class numbers and missing rates, illustrating how the transition point from LDS-dominant to LCD-dominant heterogeneity shifts across settings.

Table 7: Critical client counts K_{crit} for different class numbers and missing rates

Classes (C)	MR=0.3	MR=0.5	MR=0.7
10	420	351	150
20	688	528	260

Key Observations:

Missing Rate Effect: Higher missing rates reduce the label coverage rate $\bar{p}$ directly, leading to κ and K_{crit} reduction and making systems more vulnerable to LCD dominance at lower client counts.

Class Count Effect: K_{crit} increases proportionally with C, confirming the theoretical prediction $K_{\text{crit}} \propto C \ln C$.

For $C = 50$, the DSR also decreases as the missing rate increases, consistent with the cases $C = 10$ and $C = 20$. When $MR = 0.3$, the DSR continues to decrease over the range $K \in [10, 1000]$, although at a decelerating rate. This trend suggests that $K_{\text{crit}} > 1000$ in this regime. Because the magnitudes of C, K, and N_{total} are all substantially larger in this setting, we omit a more detailed analysis here.

Theoretical Validation: The empirical DSR trends validate our theoretical framework

$$\text{DSR} = \frac{\text{SSDI}_{\text{LCD}}}{\text{SSDI}_{\text{LDS}}} \sim \frac{O(CK^2)}{O(C^3 \ln^2 C)} = \frac{K^2}{C^2 \ln^2 C}$$

This theoretical relationship explains the U-shaped trend and provides the scaling law for critical points.

C.2.3 Critical Point Analysis for Varying Classes

Empirical Critical Point Determination: As presented in Figure 2a and Figure 2b, we determine K_{crit} experimentally by:

1. Computing DSR values across a range of client counts K.
2. Fitting a $f(x) = a \cdot x + b \cdot x^{-1}$ shape function to the DSR vs K curve.
3. Identifying the extreme point as K_{crit}.
4. Repeating for different class counts C and missing rates MR.

Missing Rate Modulation: The critical scaling relationship is modulated by the missing rate

$$K_{\text{crit}}(MR) = \kappa(MR) \cdot C \ln C,$$

where $\kappa(MR)$ decreases monotonically with increasing missing rate.

Practical Implications:

System Capacity Planning: The critical scaling provides concrete guidance for determining maximum supportable clients given class diversity.

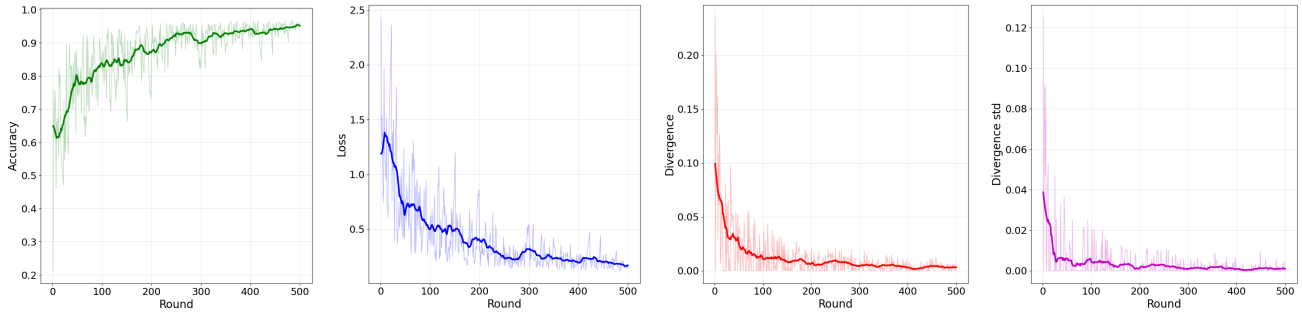
Robustness Design: Understanding critical points helps design systems that avoid LCD-dominant regimes where performance degradation is most severe(shown in C.3).

C.2.4 Performance Comparison Across LDS- and LCD-Dominated Regimes

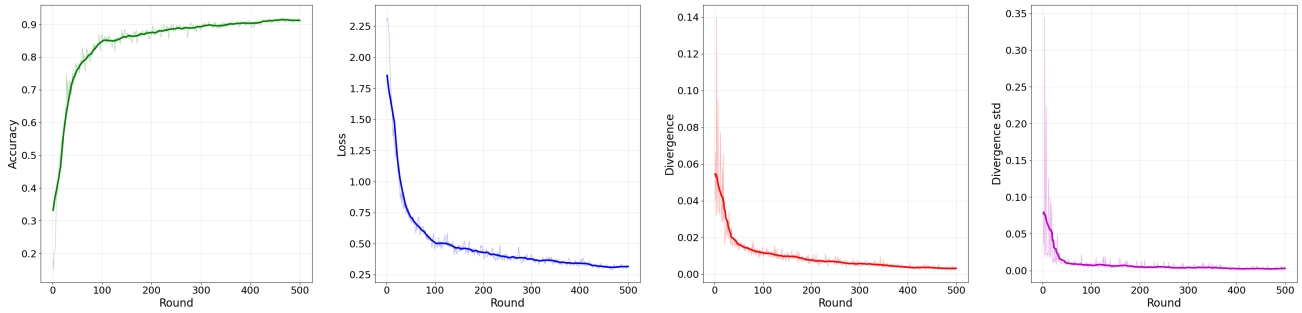
As shown in Fig. 6, we compare model performance under two client scales ($K = 20$ versus $K = 500$) on **MNIST** with $C = 10$. The smaller-client configuration ($K = 20$) achieves better final accuracy and lower loss, despite lower stability, as indicated by a larger standard deviation of client divergence and slower convergence during training.

Figure 7 reports the corresponding comparison on **CIFAR-100 (coarse)** with $C = 20$. Although the smaller-client setting again yields stronger final performance, the dynamics are more complex because CIFAR-100 is substantially harder than MNIST and is trained with ShuffleNetV2 rather than a shallow two-layer network.

Compared to the nearly perfect performance achieved on MNIST, CIFAR-100 trained with ShuffleNetV2 under a moderate heterogeneity level only attains an accuracy of 30% to 40%. Experimental observations indicate that client population size exerts a systematic influence on the training dynamics. The following summarises the key observations under two typical population regimes:

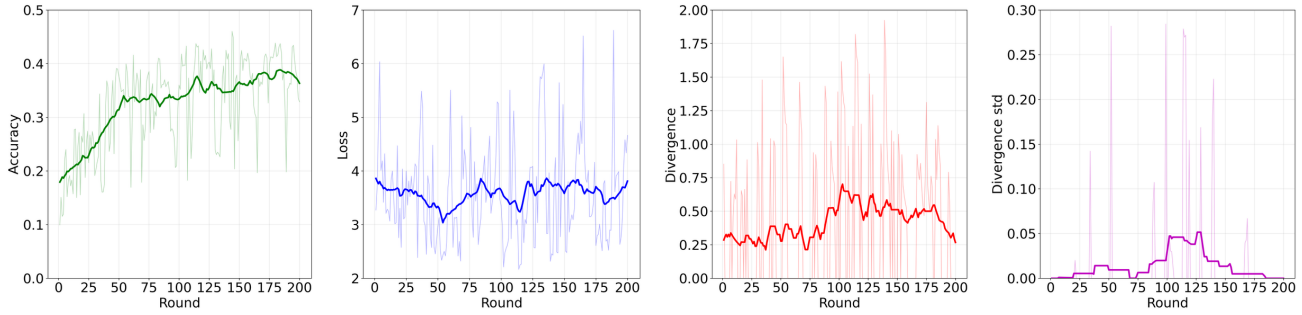


(a) Accuracy (K=20) (b) Loss (K=20) (c) Divergence (K=20) (d) Divergence Std (K=20)

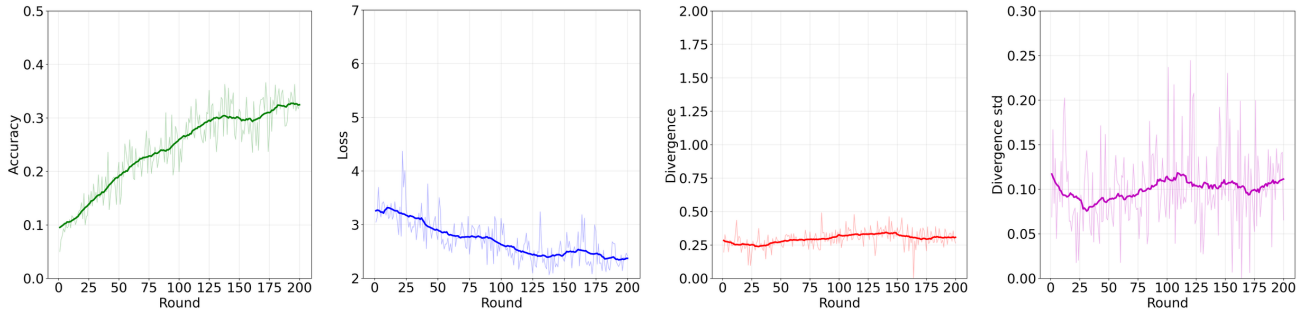


(e) Accuracy (K=500) (f) Loss (K=500) (g) Divergence (K=500) (h) Divergence Std (K=500)

Figure 6: **MNIST.** Accuracy rises rapidly before gradually plateauing, while loss, divergence, and the standard deviation of divergence decrease before levelling off. Larger K leads to more pronounced oscillations.



(a) Accuracy (K=20) (b) Loss (K=20) (c) Divergence (K=20) (d) Divergence Std (K=20)



(e) Accuracy (K=500) (f) Loss (K=500) (g) Divergence (K=500) (h) Divergence Std (K=500)

Figure 7: **CIFAR-100**. For small K, the accuracy first increases and then stabilises, while the loss first decreases, then recovers, and finally stabilizes. The divergence metric exhibits strong oscillations. The subsequent rise in divergence standard deviation exhibits a lag compared to the trends of accuracy and loss. For large K, divergence remains low and stable, whereas the divergence standard deviation remains high and oscillates intensely.

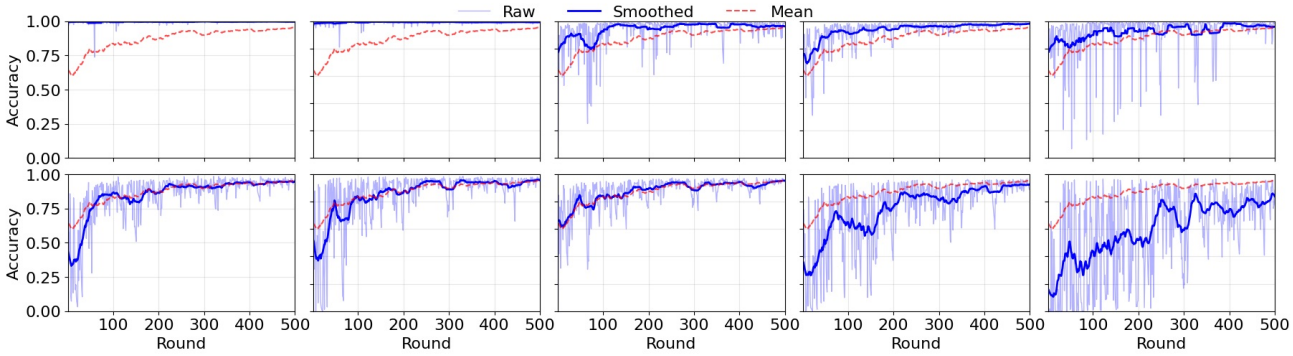
In the small-scale client regime, where the system is dominated by LDS-induced error and the total number of participating clients is limited, the training process exhibits significant instability. After initial rapid decreases in loss and increases in accuracy, both metrics begin to exhibit sustained and substantial oscillations starting from a specific training round. Concurrently, the divergence metric between client models and the global model remains at a high level and oscillates in sync with the loss and accuracy. The variance time series of this divergence metric shows sporadic extreme peaks, indicating intermittent severe conflicts in update directions across clients. Ultimately, the test accuracy of the model converges to similar levels across all classes, without significant inter-class differentiation.

In contrast, in the large-scale client regime, where the system is dominated by LCD-induced error and the number of participating clients is substantially larger, the training dynamics display distinct characteristics. The loss and accuracy metrics show slow yet monotonic improvement, eventually plateauing in the later stages of training. Notably, the value of the divergence metric between client and global models is significantly lower than in the small-scale regime and remains relatively stable throughout the training process. However, the variance of this divergence metric remains persistently high and exhibits fluctuations, reflecting sustained structural disparities among clients. As a result, the model's test accuracy shows marked differentiation across classes: higher for frequent classes and significantly lower for rare classes.

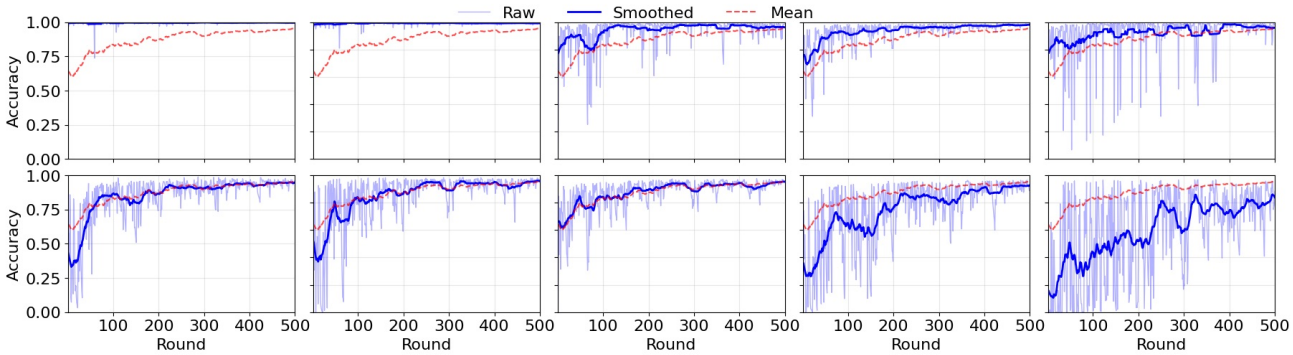
However, when using deeper architectures such as ResNet-18 and more complex, large-scale datasets like ImageNet, even with Tiny ImageNet, the system's learning capacity and resulting performance further degrade. To validate whether different regimes lead to performance divergence between frequent and rare classes, we conducted a separate per-class analysis.

C.3 Performance Divergence Patterns

C.3.1 Class-wise Accuracy Analysis

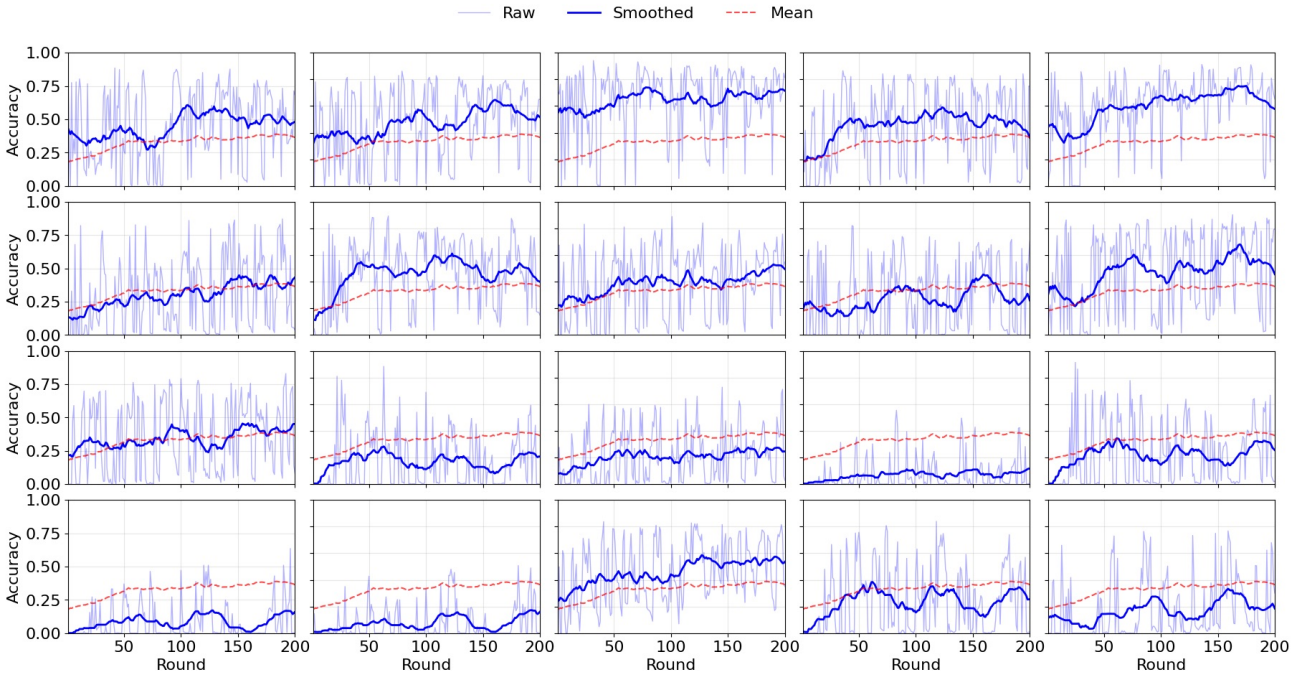


(a) MNIST, $C = 10$, $K = 20$

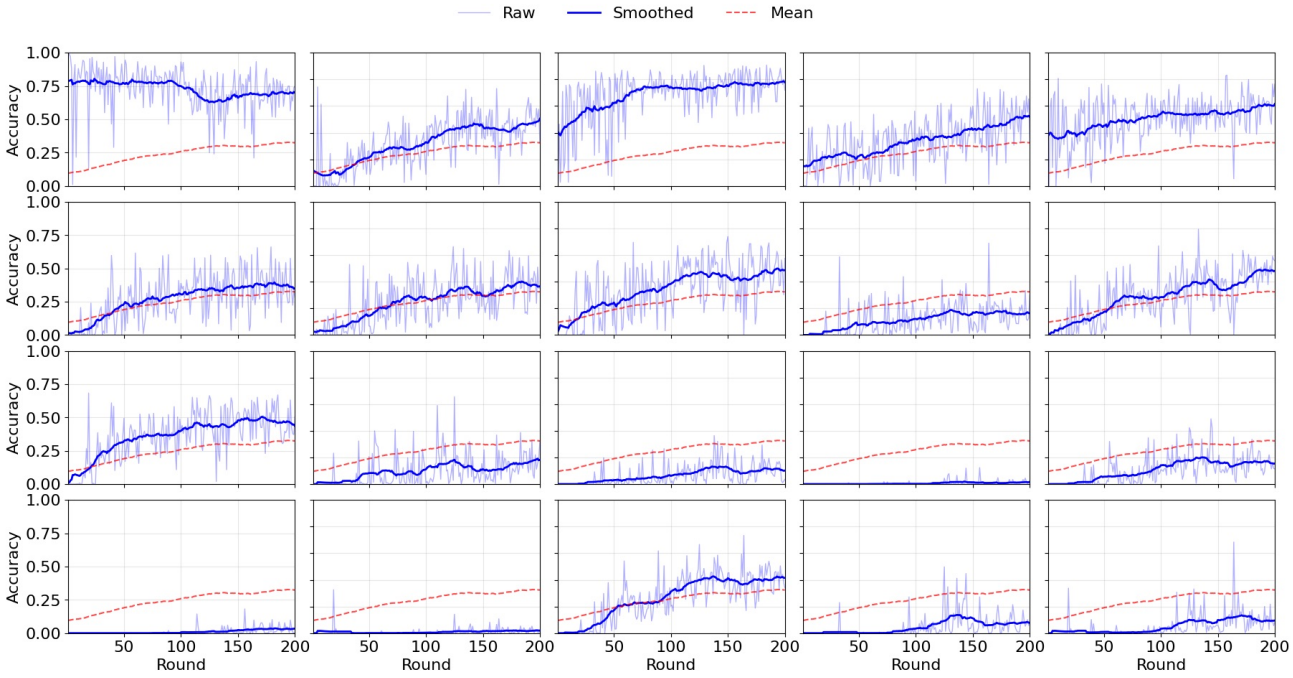


(b) MNIST, $C = 10$, $K = 500$

Figure 8: Per-class accuracy trends (part I).



(c) CIFAR-100 (coarse), $C = 20$, $K = 20$

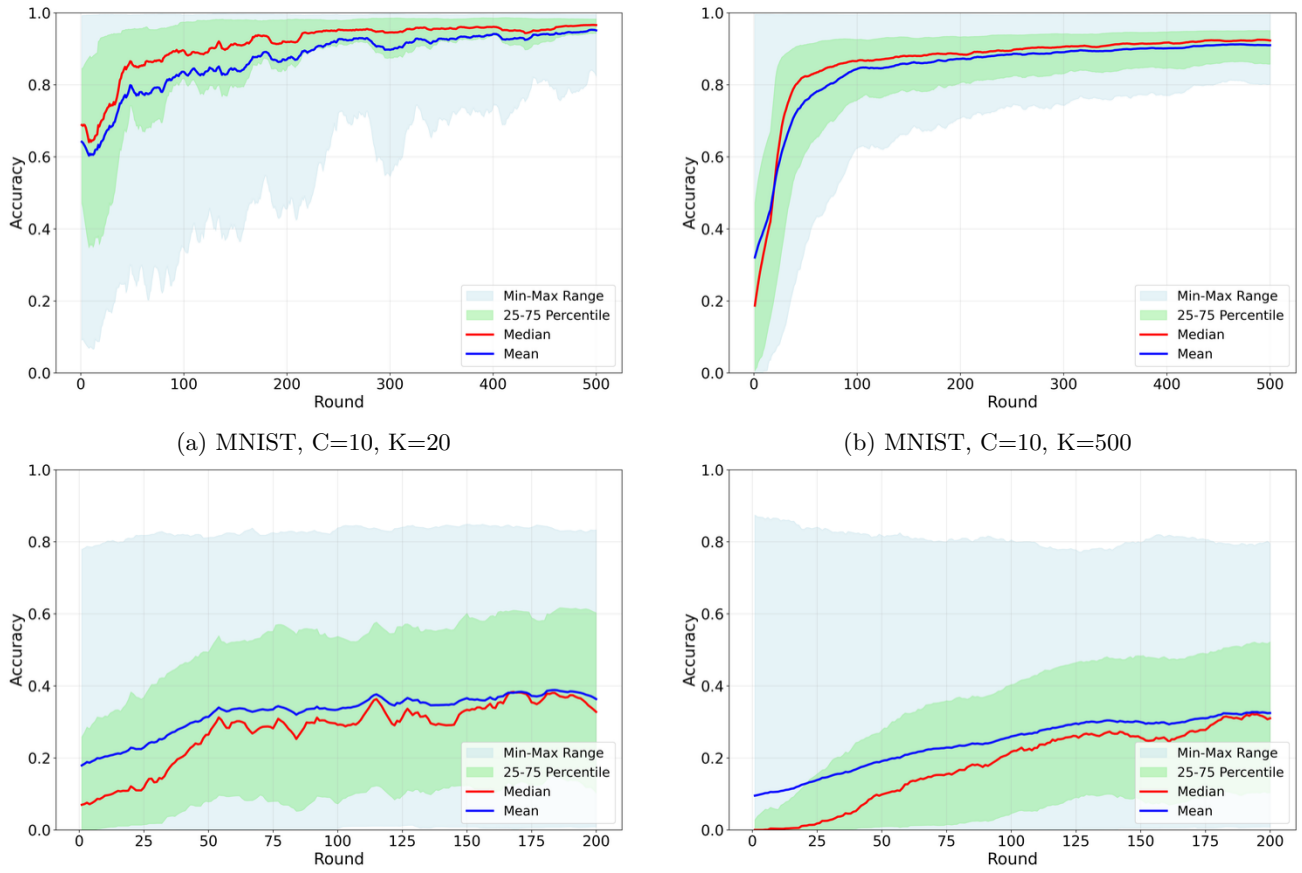


(d) CIFAR-100 (coarse), $C = 20$, $K = 500$

Figure 8: Per-class accuracy trends. (a) Frequent classes achieve near-perfect accuracy, while rare classes exhibit oscillatory and slow improvement. (b) Frequent classes approach 100% accuracy, but rare classes rapidly converge to a stable upper threshold. (c) Severe oscillations with clearly stratified accuracy levels emerge as training progresses. (d) Relatively stable trends with extreme accuracy polarisation: some labels significantly exceed average performance, while others remain near zero accuracy, indicating failure in the prediction task.

Figure 8 illustrates the class-wise accuracy trajectories under different client counts and missing rates, highlighting the growing performance disparity between frequent and rare labels as heterogeneity intensifies.

We recorded the accuracy trends of different classes throughout the training process under various labels, clients, trained models, and datasets, as illustrated in Figure 9. For MNIST, high accuracy was achieved regardless of whether the number of clients was large or small, whereas for CIFAR-100, the accuracy remained below 40%. Across all datasets, scenarios with a smaller number of clients exhibited slightly higher accuracy, yet demonstrated greater divergence among different labels and more pronounced oscillations.



(a) MNIST, C=10, K=20

(b) MNIST, C=10, K=500

(c) CIFAR100coarse, C=20, K=20

(d) CIFAR100coarse, C=20, K=500

Figure 9: Class-level accuracy trajectories during training. MNIST uses a 2-layer neural network for 500 communication rounds. CIFAR-100 (coarse) uses ShuffleNetV2 for 200 communication rounds.

A systematic shift in the dominant mechanism:

A. **Small-scale client regime: dynamic instability driven by skewed label distributions**

In this regime, the dynamic instability of the global model primarily stems from conflicting update directions induced by the skewed label distribution within individual clients.

Due to the limited total number of clients, the subset of clients participating in each communication round constitutes a biased sampling of the global distribution. Each client performs local training based on its own skewed label distribution, resulting in model update vectors that are strongly biased towards its locally dominant classes.

When the dominant classes of the sampled client subsets differ significantly across consecutive rounds (e.g., one round is dominated by clients rich in class A, the next by clients rich in class B), the model aggregation operation on the server receives a series of opposing update vectors. This causes the global model parameters W_{global} to oscillate periodically within the parameter space, rather than converging along a consistent direction.

Phenomenon Interpretation:

1. High and Oscillating Divergence:
   Directly reflects the drastic shifts in the aggregated model's direction between consecutive rounds.
2. Sparse Peaks in Divergence Variance:
   The peaks correspond to rounds where client subsets with extremely specialised data distributions (i.e., featuring highly dominant classes) were sampled. The update directions in these rounds deviate substantially from those in typical rounds.
3. Convergence to Mediocre Class Accuracy:
   As the model oscillates, it is forcibly exposed to different skewed versions of all classes. Its parameter updates are effectively averaged over time, leading to a homogenised discriminative ability across all classes, resulting in a mediocre performance level.

B. **Large-scale client regime: structural performance limitations driven by label absence**

In this regime, the training dynamics are dominated by the prevalent label absence across clients.

The large number of clients leads to the global dataset being partitioned into numerous data islands, each containing only a small subset of classes. For any given class A in the global class set, its gradient signal is provided solely by the subset of clients that contain A.

Pseudo-Consistent Updates: Because a large number of clients participate each round, their collective data distribution approximates a stable, albeit biased, global prior. High-frequency classes, present in many clients, have their gradients reliably and stably provided during each aggregation, enabling their continuous optimisation.

Gradient Signal Dilution: Low-frequency classes exist only within very few clients. During aggregation, gradients corresponding to these rare classes, provided by these few clients, become diluted amidst the vast number of updates from other clients. Consequently, the global model fails to effectively adjust the decision boundaries for low-frequency classes.

Phenomenon Interpretation:

1. Low and Stable Divergence:
   Each aggregation is based on a stable sampling of the client distribution, resulting in consistent update directions. Consequently, the average divergence between client models and the global model is reduced and stabilised.
2. Persistently High Divergence Variance:
   Despite the low average divergence, the level of divergence individual clients exhibit is highly polarised. Clients containing common classes show low divergence, whereas those acting as data islands containing rare classes exhibit extremely high divergence. The oscillating high variance of divergence reflects the randomness in sampling these high-divergence clients in each communication round.
3. Polarised Accuracy and Plateau:
   The model's performance ceiling is determined by the learnable high-frequency classes. Once the accuracy for these classes saturates, the overall performance plateaus because the low-frequency classes cannot be effectively learned due to missing gradient signals.

This class-level analysis reveals that in highly heterogeneous federated learning environments, the client population size is a critical factor determining the training dynamics and final model performance properties. The core insight is that the change in scale induces a shift in the dominant challenging mechanism:

In the small-scale client regime, the core challenge is the dynamic instability caused by **skewed label distribution**, manifesting as oscillations in the training process.

In the large-scale client regime, the core challenge shifts to the structural performance limitation caused by **label absence**, manifesting as model bias and an insurmountable performance ceiling.

C.3.2 Rare vs Frequent Label Performance

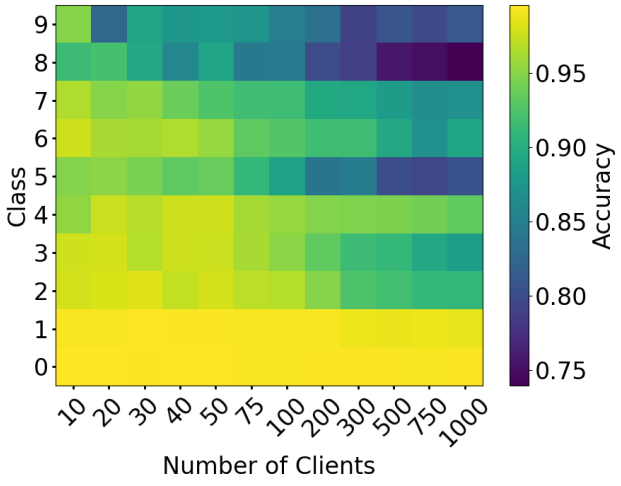
Figure 10 shows both the accuracy trajectories of rare and frequent labels and their corresponding performance gaps under different missing rates, highlighting the widening disparity as the client count increases.

We consider client counts $K \in \{10, 20, 30, 50, 75, 100, 200, 300, 500, 750, 1000\}$ and two representative missing rates, $MR \in \{0.3, 0.5\}$, while fixing $C \in \{10, 20\}$. Labels are grouped according to their global prevalence under the Zipf distribution:

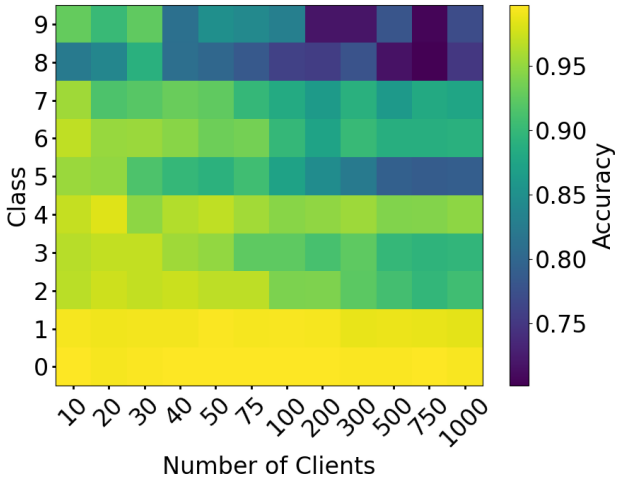
Frequent Labels: Top 20% of classes by global sample count (classes 0–1 for $C = 10$).

Rare Labels: Bottom 20% of classes by global sample count (classes 8–9 for $C = 10$).

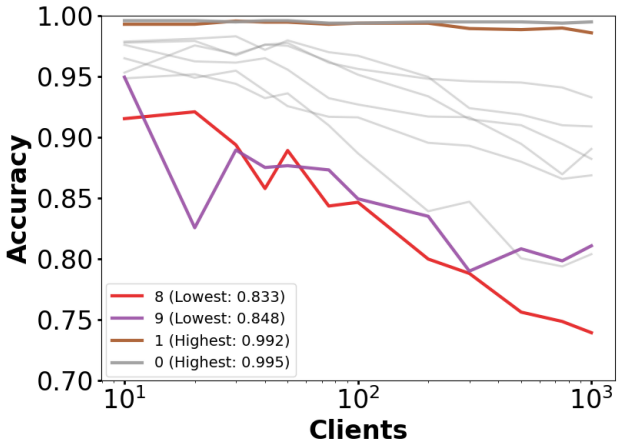
Intermediate Labels: Remaining 60% of classes showing transitional behaviour.



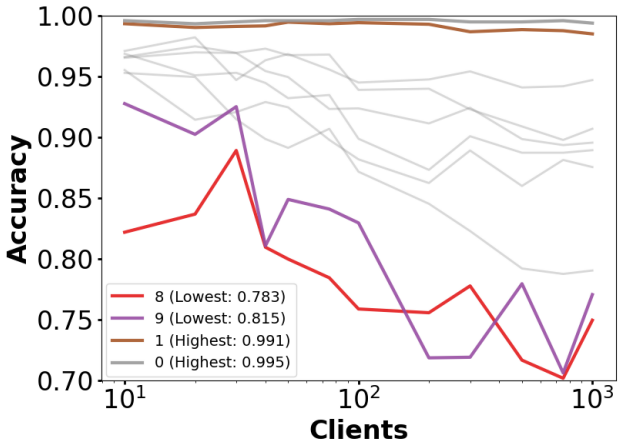
(a) MNIST, accuracy, $MR = 0.3$



(b) MNIST, accuracy, $MR = 0.5$



(c) MNIST, gap, $MR = 0.3$



(d) MNIST, gap, $MR = 0.5$

Figure 10: Accuracy trajectories of frequent and rare labels (top row) and their performance gaps (bottom row) under different missing rates. As the client count increases, rare labels degrade earlier and more severely than frequent labels, and this disparity becomes more pronounced under higher missing rates.

Performance Divergence and Collapse Patterns:

A. **Stability boundary:** Frequent labels maintain comparatively high accuracy (MNIST: ~ 0.95, CIFAR-100: ~ 0.65, Tiny ImageNet: ~ 0.40).

B. **Nonlinear decay:** Rare-label accuracy decays nonlinearly, with faster deterioration once K exceeds the critical regime.

C. **Increasing disparity:** The accuracy gap between rare and frequent labels widens as K increases, together with the variance across classes.

D. **Missing-rate effect:** Higher missing rates induce earlier and more severe divergence in class-level accuracy.

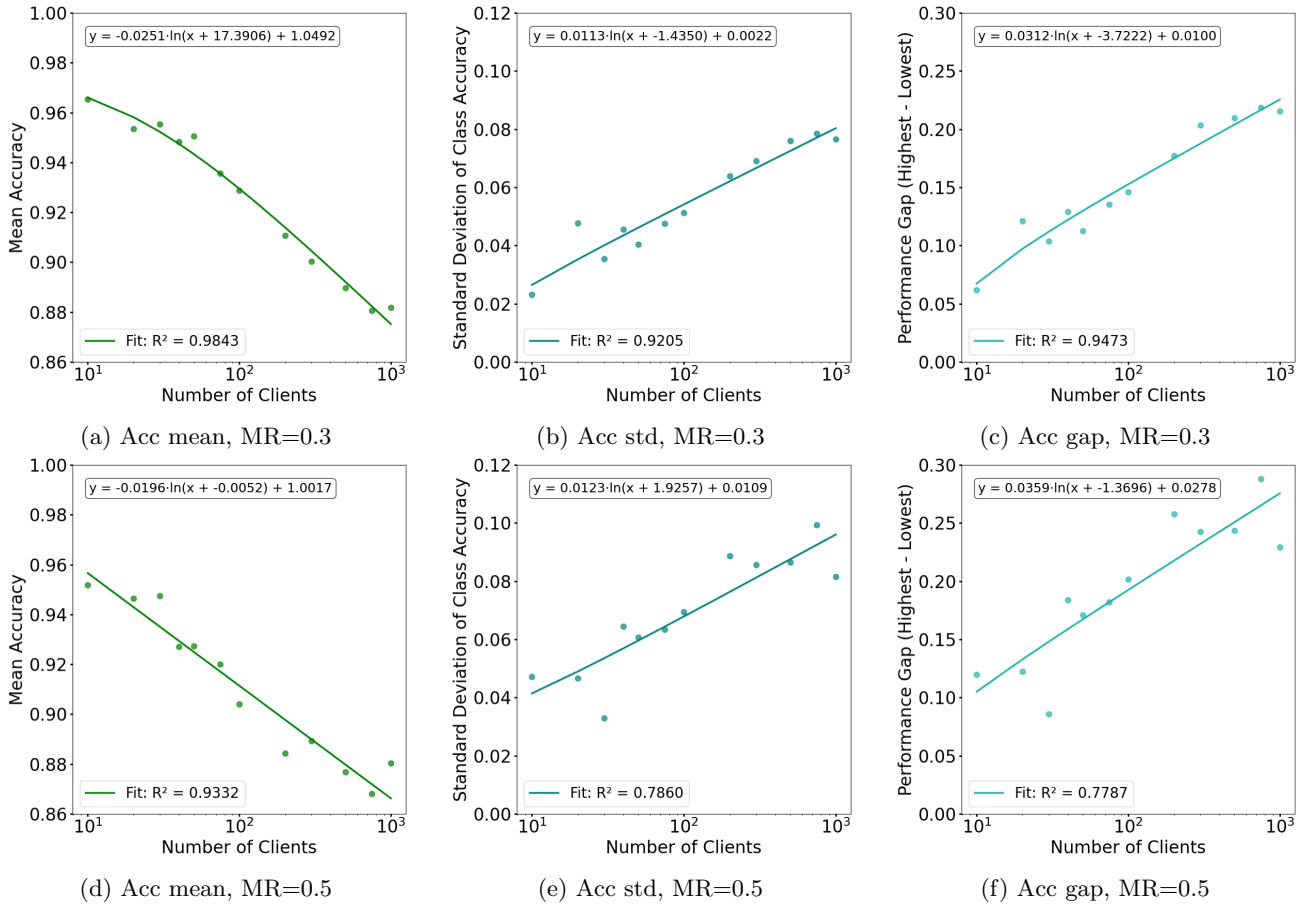
E. **Prediction failure:** When K becomes sufficiently large, some classes remain near zero accuracy; the number of such failed classes increases with K.

Theoretical Explanation: The divergence between rare and frequent labels follows directly from our theoretical framework: rare labels suffer more severely from coverage deficiency due to their lower η_c in Term D of Γ_{LDS} (LDS Penalty), experience fewer learning opportunities as a result of smaller N_c in both LDS and LCD penalties (Sampling Effects), and undergo a catastrophic collapse when the system transitions into the LCD-dominant phase ($K > K_{\text{crit}}$, Critical Transition).

System Design Implications: Monitoring rare label performance serves as an effective early warning mechanism for detecting heterogeneity risks; consequently, such labels necessitate targeted mitigation strategies, such as data augmentation or personalised approaches, while system scale should be planned with caution to avoid operational regimes where rare label performance collapses.

C.3.3 Accuracy Variance and Range Metrics

The relevant statistical data demonstrates a close relationship between the number of clients and class-level accuracy in Figure 11. We calculated the mean, variance, and range (gap) under different missing rates.



(a) Acc mean, MR=0.3 (b) Acc std, MR=0.3 (c) Acc gap, MR=0.3

(d) Acc mean, MR=0.5 (e) Acc std, MR=0.5 (f) Acc gap, MR=0.5

Figure 11: Mean, standard deviation, and range of class-level accuracy show strong correlations with increasing K. A higher missing rate exacerbates the disparity in accuracy.

Key Relationships: The increase in both variance and range reflects the progressive intensification of the label missing rate. While skewed label distribution may contribute to a certain level of variance, it still allows for some learning opportunities. In contrast, label absence directly renders certain classes unlearnable on affected clients, resulting in significantly elevated variance and range. As for the mean accuracy, the catastrophic failure

caused by label absence—where accuracy drops close to zero—is far more detrimental than the impact of client drift under LDS-dominated regimes, which only leads to moderately low accuracy. Moreover, an increase in the missing rate invariably intensifies heterogeneity, thereby exacerbating class-level divergence and causing these effects to manifest more severely and at a smaller critical threshold K.

Practical Applications: The variance and range of class-level accuracy serve as straightforward and interpretable metrics for real-time system health monitoring. Furthermore, thresholds defined on these metrics can effectively trigger adaptive mitigation strategies when necessary. These standardised metrics also facilitate consistent benchmarking of heterogeneity effects across diverse systems and algorithms. Importantly, elevated variance and range values signal potential fairness concerns that warrant further investigation and intervention.

C.4 Robustness Analysis

We evaluate the robustness of these patterns across datasets, model architectures, and training protocols:

Different datasets: Consistent qualitative patterns are observed across MNIST, CIFAR-100, and Tiny ImageNet subsets.

Model architectures: Similar divergence patterns are observed for both ResNet-18 and ShuffleNetV2.

Training protocols: The main trends persist across different random initialisations and data allocations.

As we have previously validated across different datasets and model architectures, although datasets introduce significant variations in absolute accuracy, the underlying outcomes consistently adhere to a shared set of patterns, independent of dataset complexity. The choice of model, particularly more complex ones such as ResNet-18, may slow or even obscure certain performance dynamics. However, when analysed in conjunction with metrics such as loss and divergence, these patterns remain traceable and interpretable.

All prior experiments were conducted under strictly controlled conditions with fixed parameters including C and K. However, the influence of total data size N, representing an $O(1)$-order variation, has not been systematically examined previously. This section is therefore dedicated to analysing the effects of varying N.

C.4.1 Impact of Total Dataset Size on SSDI

Figure 12 presents the robustness of accuracy, loss, divergence, and divergence variability across different total dataset sizes, allowing us to evaluate whether the observed SSDI-induced performance patterns remain stable when the total number of samples varies while the heterogeneity structure is kept comparable.

Experimental Design: We generated a series of distinct label-client matrix files for each dataset sample size, while ensuring the total sample volume remained within the range satisfying $C \times K$. The comparative results of these matrix files have been presented in the main text. Under identical data distribution patterns (Zipf and Pareto), even randomly generated matrices exhibit SSDI fluctuations within a narrow range. We utilised this series of matrix files to guide the generation of corresponding datasets for training, collected the results, and analysed metrics including accuracy, loss, and divergence.

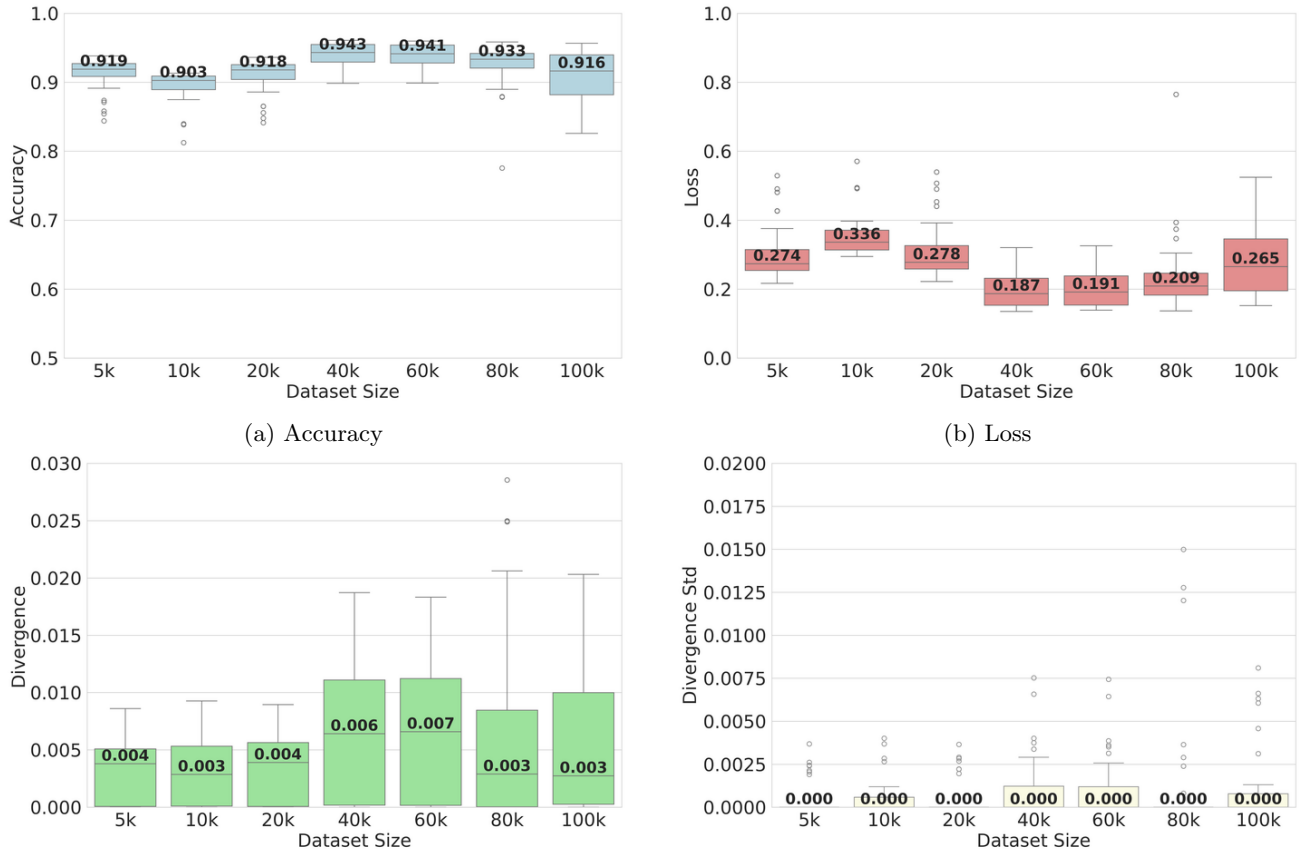
We systematically varied the N_{total} across [3k, 5k, 10k, 20k, 40k, 60k, 80k, 100k] while maintaining fixed heterogeneity structure, with constant label missing rates of 0.3 and 0.5, data distribution parameters Zipf=1.0 and Pareto=2.0, and comparable levels of label absence skewness. The models were trained across different combinations of [C, K] and experimental data were collected for analysis.

Theoretical Explanation: The stability of SSDI with respect to dataset size follows from its mathematical formulation:

- Normalisation Property: SSDI is defined as a ratio of normalised deviations:

$$\text{SSDI} = \frac{\|W \odot D\|_F}{V_{\max}(C, K)}$$

  Both numerator and denominator scale proportionally with data volume when heterogeneity structure is fixed.



(a) Accuracy

(b) Loss

(c) Divergence

(d) Divergence standard deviation

Figure 12: Model performance remains broadly stable across different total dataset sizes.

- Distribution Preservation: Under fixed distribution parameters (ρ_c^{global}, n_k/N_{total}), the deviation matrix D scales linearly while the (C, K)-adaptive normalization constant $V_{\max}(C, K)$ remains fixed, thereby preserving SSDI.

- Asymptotic Independence: Our theoretical analysis predicted $\Gamma_{\text{LDS}} \sim O(C^3 \ln^2 C)$ and $\Gamma_{\text{LCD}} \sim O(CK^2)$, both independent of lower order term N_{total}.

Practical Implications: The SSDI metric demonstrates its practical value through its scale invariance, providing consistent heterogeneity measurements across diverse data regimes from small-scale prototypes to large-scale deployments. This structural focus enables the metric to capture distribution patterns rather than being influenced by absolute data volume, making it particularly suitable for cross-study comparisons. Additionally, the framework supports reliable heterogeneity assessment even with limited data samples, facilitating early-stage system evaluation and intervention.

C.4.2 Scalability with Client and Class Numbers

Computational Complexity Analysis:

We analyse SSDI scalability through both theoretical analysis and empirical measurements:

Theoretical Complexity: SSDI calculation requires $O(CK)$ operations for matrix constructions and norm calculations.

Memory Requirements: Storage of $C \times K$ matrices requires $O(CK)$ memory.

Comparison with Alternative Metrics:

Table 8 compares the computational and memory complexities of SSDI with several commonly used heterogeneity-related baselines, highlighting the linear scalability of SSDI with respect to C and K.

Table 8: Computational comparison with baseline heterogeneity metrics

Metric	Time Complexity	Space Complexity
SSDI (Ours)	$O(CK)$	$O(CK)$
Earth Mover's Distance	$O(C^2K^2)$	$O(CK)$
Dirichlet-α	$O(CK)$	$O(C+K)$
Performance Variance	$O(max(C,K))$	$O(C)$

Cross-Dataset Consistency:

Table 9 summarises the per-class accuracy statistics across datasets, showing that increasing task complexity leads to both lower mean accuracy and substantially larger class-level disparity.

Table 9: Per-class Accuracy Statistics under Different Datasets (K=75,MR=0.5)

Dataset	Label	Max Acc	Min Acc	Mean Acc	Std Acc	Median Acc
MNIST	C=10	0.9959	0.8840	0.9593	0.0382	0.9752
CIFAR-100	C=20	0.8120	0.0000	0.3288	0.3200	0.2250
Tiny ImageNet	C=50	0.7800	0.0000	0.1708	0.1798	0.1200

C.5 Implementation and Reproducibility

C.5.1 Computational Infrastructure

Tables 10 and 11 summarise the hardware and software environment used to ensure the reproducibility of the reported experiments.

Table 10: Hardware environment used in the experiments.

Component	Specification
Computing Platform	Google Colaboratory
Accelerator	NVIDIA Tesla T4 (16GB)
CPU	Intel Xeon CPU @ 2.20GHz (2 cores)
Memory	12.7 GB RAM
Storage	108GB total (69GB available)
Python Version	Python 3.12.11

Table 11: Software environment used in the experiments.

Component	Version/Description
Deep Learning Framework	PyTorch 2.8.0+cu126
Federated Learning Implementation	Custom framework built on PyTorch
Python Version	3.12.11
Key Dependencies	NumPy 2.0.2, pandas 2.2.2, SciPy 1.16.2, Matplotlib 3.10.0
Environment	Google Colaboratory notebook environment

C.5.2 SSDI Calculation Algorithm

Algorithm 2 summarises the full computation pipeline of SSDI and its LDS/LCD components from the client-class count matrix.

Algorithm 2 SSDI Calculation from Client-Class Counts

Input: Client-class counts $\mathbf{N} = [n_{k,c}]$, $k = 1, \ldots, K$, $c = 1, \ldots, C$
Output: SSDI, SSDI_{LDS}, SSDI_{LCD}, DSR

Step 1. Basic marginals
$N_{\text{total}} \leftarrow \sum_{k=1}^{K} \sum_{c=1}^{C} n_{k,c}$ ▷ total samples
$N_c \leftarrow \sum_{k=1}^{K} n_{k,c}, \ \forall c$ ▷ global class counts
$n_k \leftarrow \sum_{c=1}^{C} n_{k,c}, \ \forall k$ ▷ client sample counts
$\rho_c^{\text{global}} \leftarrow N_c / N_{\text{total}}, \ \forall c$ ▷ global class proportions
$\rho_{k,c} \leftarrow n_{k,c} / n_k, \ \forall (k, c)$ ▷ local class proportions

Step 2. Fine-grained deviations
for $c = 1$ to C **do**
  $d^c \leftarrow \sqrt{\frac{1}{N_c N_{\text{total}}} \sum_{k=1}^{K} n_k^2 \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}$ ▷ class-level LCD, Eq. (1)
end for
for $k = 1$ to K **do**
  $d_k \leftarrow \sqrt{\frac{n_k}{N_{\text{total}}} \sum_{c=1}^{C} \left(\rho_{k,c} - \rho_c^{\text{global}}\right)^2}$ ▷ client-level LDS, Eq. (2)
end for
$\mathbf{v} \leftarrow (d^1, \ldots, d^C, d_1, \ldots, d_K)^\top$ ▷ unified deviation vector, Eq. (3)

Step 3. Matrix form
for each (c, k) **do**
  $p_{c,k} \leftarrow n_{k,c} / N_{\text{total}}, \quad q_{c,k} \leftarrow \rho_c^{\text{global}} \, n_k / N_{\text{total}}$ ▷ joint and IID baseline
  $d_{c,k} \leftarrow p_{c,k} - q_{c,k}, \quad w_{c,k} \leftarrow \sqrt{N_{\text{total}}/N_c + N_{\text{total}}/n_k}$ ▷ deviation and weight
end for

Step 4. LDS/LCD decomposition
for each (c, k) **do**
  if $n_{k,c} > 0$ **then**
    $d_{c,k}^{\text{LDS}} \leftarrow d_{c,k}, \quad d_{c,k}^{\text{LCD}} \leftarrow 0$ ▷ present labels
  else
    $d_{c,k}^{\text{LDS}} \leftarrow 0, \quad d_{c,k}^{\text{LCD}} \leftarrow -q_{c,k}$ ▷ missing labels
  end if
end for

Step 5. Normalization
$\text{num} \leftarrow \|W \odot D\|_F, \quad \text{num}_{\text{LDS}} \leftarrow \|W \odot D^{\text{LDS}}\|_F, \quad \text{num}_{\text{LCD}} \leftarrow \|W \odot D^{\text{LCD}}\|_F$ ▷ weighted magnitudes
$m \leftarrow \min(C, K)$
$\text{den} \leftarrow V_{\max}(C, K) = \sqrt{2\left(1 - \frac{1}{m}\right)}$ ▷ (C, K)-adaptive theoretical normalization

Step 6. Final scores
$\text{SSDI} \leftarrow \text{num}/\text{den}$ ▷ overall heterogeneity
$\text{SSDI}_{\text{LDS}} \leftarrow \text{num}_{\text{LDS}}/\text{den}, \quad \text{SSDI}_{\text{LCD}} \leftarrow \text{num}_{\text{LCD}}/\text{den}$ ▷ component scores
$\text{DSR} \leftarrow \text{SSDI}_{\text{LCD}}/\text{SSDI}_{\text{LDS}}$ ▷ deficiency-to-skew ratio
return SSDI, SSDI_{LDS}, SSDI_{LCD}, DSR

Algorithm Properties:

Time complexity: $O(CK)$ operations are required for the marginal computations, matrix construction, and norm evaluation.

Space complexity: $O(CK)$ memory is required to store the $C \times K$ matrices P, Q, D, and W.

Numerical stability: The implementation guards against division by zero and numerical underflow.

Parallelisation: Matrix operations can be parallelised to improve efficiency.

C.6 Algorithm Sensitivity under Structured SSDI Regimes

In this appendix, we provide additional empirical analysis to illustrate how different federated optimization algorithms behave under controlled heterogeneity regimes characterized by SSDI. Rather than focusing on individual heterogeneity realizations, we evaluate algorithm sensitivity across multiple structured data partitions generated with predefined SSDI levels. This allows us to examine how algorithmic behavior evolves as heterogeneity increases.

Experimental protocol: Datasets are generated using the structured data generation procedure described in the main paper. For each target SSDI level, multiple independent heterogeneity realizations are generated, and the reported results correspond to the mean and standard deviation across these realizations. All algorithms are trained using identical model architectures and training hyperparameters.

Table 12: Average test accuracy (mean ± std) over structured partitions under different SSDI regimes.

K	SSDI	FedAvg	FedProx	SCAFFOLD	Ditto
20	0.2	0.966±0.005	0.966±0.005	**0.971±0.002**	0.966±0.006
20	0.4	0.889±0.046	0.889±0.044	**0.909±0.038**	0.888±0.043
20	0.6	0.353±0.040	**0.354±0.040**	0.281±0.075	0.352±0.039
20	0.8	**0.199±0.041**	0.195±0.037	0.153±0.035	0.197±0.041
300	0.2	0.908±0.003	0.909±0.003	**0.915±0.002**	0.909±0.002
300	0.4	0.708±0.036	0.705±0.036	**0.736±0.056**	0.701±0.039
300	0.6	0.326±0.069	0.325±0.069	**0.390±0.067**	0.327±0.068
300	0.8	0.104±0.015	0.108±0.017	**0.241±0.098**	0.108±0.019

Table 13: Worst-class accuracy (mean ± std) under structured SSDI regimes.

K	SSDI	FedAvg	FedProx	SCAFFOLD	Ditto
20	0.2	0.921±0.044	0.923±0.042	**0.945±0.008**	0.919±0.046
20	0.4	0.624±0.245	0.624±0.245	**0.655±0.222**	0.628±0.239
20	0.6	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
20	0.8	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
300	0.2	0.816±0.021	0.819±0.022	**0.828±0.018**	0.821±0.017
300	0.4	0.068±0.151	0.046±0.103	**0.075±0.169**	0.064±0.143
300	0.6	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
300	0.8	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000

Table 14: Standard deviation of class-wise accuracy (mean ± std) under structured SSDI regimes.

K	SSDI	FedAvg	FedProx	SCAFFOLD	Ditto
20	0.2	0.001±0.001	0.000±0.000	0.000±0.000	**0.001±0.001**
20	0.4	**0.018±0.023**	0.018±0.022	0.015±0.019	0.018±0.022
20	0.6	0.169±0.033	**0.170±0.032**	0.129±0.048	0.168±0.034
20	0.8	**0.086±0.021**	0.084±0.022	0.080±0.028	0.085±0.022
300	0.2	**0.003±0.001**	0.003±0.001	0.003±0.001	0.003±0.001
300	0.4	0.093±0.027	**0.097±0.025**	0.090±0.034	0.097±0.028
300	0.6	0.171±0.014	0.171±0.014	**0.185±0.007**	0.172±0.014
300	0.8	0.090±0.000	0.089±0.001	**0.136±0.046**	0.089±0.001

We evaluate four widely used federated learning algorithms: FedAvg, FedProx, SCAFFOLD, and Ditto. All experiments are conducted on MNIST with a two-layer neural network (MNIST-2NN). We consider two representative

client scales ($K = 20$ and $K = 300$) and four SSDI regimes ($0.2, 0.4, 0.6$, and 0.8).

Table 12 reports the average test accuracy under different SSDI regimes. To better understand tail behavior and class imbalance effects, Table 13 reports the worst-class accuracy, and Table 14 reports the standard deviation of class-wise accuracy.

Observations and implications: Several consistent patterns emerge from these results.

First, overall test accuracy degrades monotonically as SSDI increases (Table 12), confirming that larger SSDI values correspond to more severe heterogeneity regimes.

Second, tail-class performance deteriorates much faster than average accuracy. As shown in Table 13, worst-class accuracy collapses to nearly zero when SSDI reaches 0.6 or above for both client scales. This indicates that coverage deficiency becomes the dominant factor governing tail performance in high-heterogeneity regimes.

Third, algorithmic differences become more visible under moderate heterogeneity levels. For example, SCAFFOLD achieves slightly higher average accuracy when SSDI is low to moderate (e.g., SSDI = 0.2 or 0.4), while its advantage diminishes as heterogeneity increases.

Finally, increasing the number of clients from $K = 20$ to $K = 300$ does not eliminate tail degradation under high SSDI regimes. Although the overall heterogeneity structure changes, the collapse of worst-class accuracy persists, highlighting that coverage-related heterogeneity remains a fundamental challenge for federated learning systems.