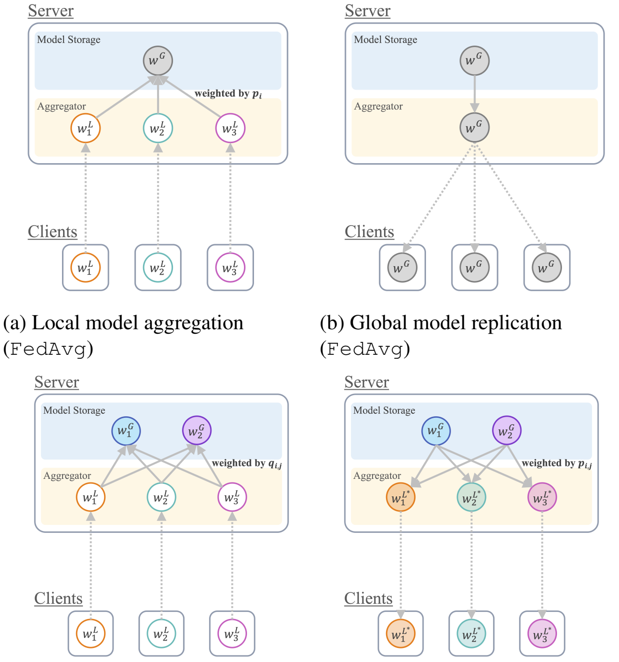
Class-Wise Federated Averaging for Efficient Personalization

Supplementary Material

A. Comparison of Aggregation Process

Figure 7 illustrates the aggregation processes of FedAvg and cwFedAvg using three clients for a binary classification task: (a) FedAvg: Server aggregates received local models. (b) Server distributes the aggregated global model to all clients. (c) cwFedAvg: Server performs class-wise aggregation to create class-specific global models. (d) Server creates personalized models by combining class-specific global models and distributes them to clients.



(a) Local model aggregation (FedAvg)

(b) Global model replication (FedAvg)

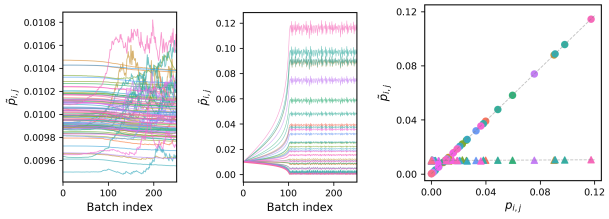
(c) Class-wise local model aggregation (cwFedAvg)

(d) Class-wise global model aggregation (cwFedAvg)

Figure 7. Comparison of aggregation processes in FedAvg and cwFedAvg with three clients for binary classification task. $*$ denotes the local models updated using class-specific global models.

B. Effect of WDR for Many-Class and Highly Imbalanced Data

Figure 8 exhibits similar patterns to the CIFAR-10 pathological setting (Figure 2). For client ID 11, which contains approximately ten dominant classes, WDR achieves better class separation (Figure 8b), resulting in $\tilde{p}_{i,j}$ values that closely match $p_{i,j}$ (circular markers in Figure 8c).



(a) Train w/o WDR (b) Train w/ WDR (c) Correlation

Figure 8. Evolution of $\tilde{p}_{i,j}$ and its correlation with $p_{i,j}$ for CIFAR-100 practical setting.

C. Experimental Details

C.1. CNN Architecture and Hyperparameters

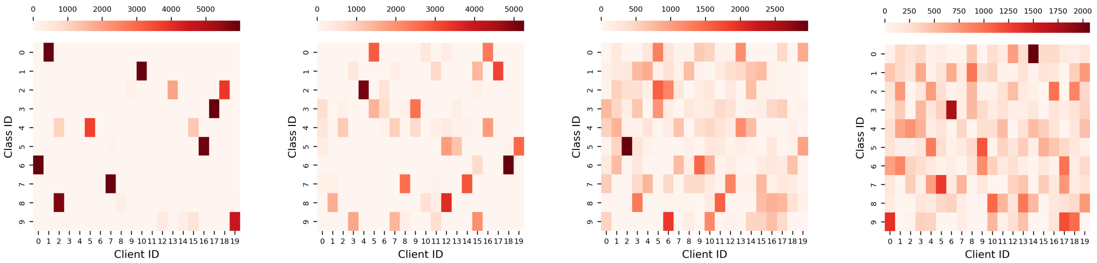
We employ a 4-layer CNN architecture [18] composed of two convolutional layers with 5×5 kernels (32 and 64 channels respectively), each paired with 2×2 max pooling. The network terminates with a fully connected layer containing 512 units and ReLU activation, followed by a softmax output layer. We adopt the hyperparameter settings from Zhang et al. [30] for baseline algorithms except for CFL, IFCA, FedNH and FedUV. For CFL, IFCA, FedNH and FedUV we follow the configurations specified in their respective papers. A comprehensive list of hyperparameter settings for all baselines is provided in Table 5.

Algorithm	Hyperparameter settings
FedProx	μ (proximal term) $= 0.001$
FedAMP	α_k (gradient descent) $= 1000$ λ (regularization) $= 1$ σ (attention-inducing function) $= 0.1$
CFL	ϵ_1 (norm of averaged updated weight) $= 0.4$ ϵ_2 (norm of maximum updated weight) $= 0.9$
IFCA	k (number of clusters) $= 2$ for CIFAR-10, 8 for CIFAR-100 and Tiny ImageNet
FedNH	ρ (smoothing parameter) $= 0.9$
FedUV	μ (classifier variance regularizer) $= 2.5$ λ (Hyperspherical uniformity regularizer) $= 0.5$

Table 5. Hyperparameter settings for the baselines.

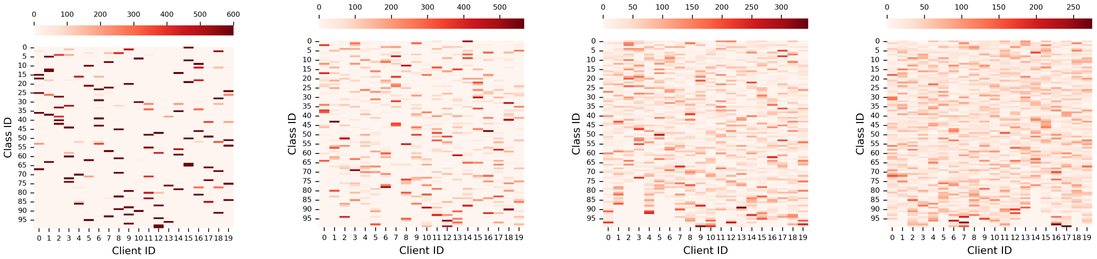
C.2. Implementation Details

The experiments are implemented in PyTorch 2.4 and conducted on a server with two Intel Xeon Gold 6240R CPUs (96 cores total), 256GB memory, and two NVIDIA RTX A6000 GPUs running Ubuntu 22.04 LTS.



(a) $\alpha = 0.01$ (b) $\alpha = 0.1$ (c) $\alpha = 0.5$ (d) $\alpha = 1.0$

Figure 9. Data distributions for the CIFAR-10 practical heterogeneous setting.



(a) $\alpha = 0.01$ (b) $\alpha = 0.1$ (c) $\alpha = 0.5$ (d) $\alpha = 1.0$

Figure 10. Data distributions for the CIFAR-100 practical heterogeneous setting.

C.3. Data Distributions of Practical Settings

Figures 9 and 10 show data distributions that vary according to different α values. Each cell in the heatmaps indicates the number of samples per class for each client. Increasing α results in decreased data heterogeneity.

D. Text Dataset Evaluation

We evaluated our approach on a text dataset to test its effectiveness across various modalities. Table 6 shows the results for the four highest-performing algorithms, reporting test accuracy on AGNews using FastText.

FedAvg	FedAMP	FedFomo	cwFedAvg
79.57 ± 0.17	97.95 ± 0.05	97.93 ± 0.09	**98.19 ± 0.01**

Table 6. Classification accuracy(%) for AG News.

E. Memory Cost Comparison

The memory efficiency of cwFedAvg depends on the ratio of parameter counts in the feature extractor to those in the classifier. Table 7 demonstrates the memory cost (number of parameters in millions) differences for ResNet-18 (512 feature dimension in the penultimate layer) with varying class counts. While increasing class counts requires higher memory costs, our selective approach (cwFedAvg (Output)) significantly reduces cost compared to the non-selective approach (cwFedAvg (All)).

# Classes	FedAvg	cwFedAvg (All)	cwFedAvg (Output)
10	11.18	111.81	**11.23**
100	11.23	1122.67	**16.36**
1000	11.69	11688.42	**524.68**

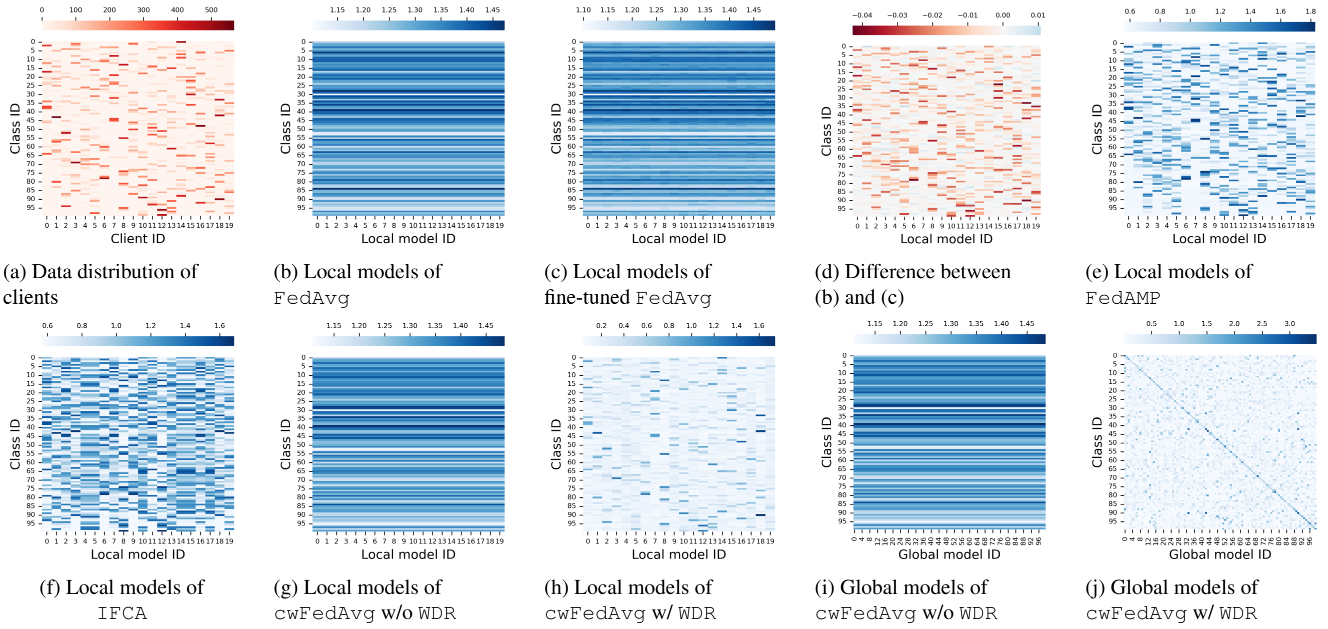
Table 7. Memory cost comparison for ResNet-18.

F. Visualizations of ℓ_2-norms of Output Layer Weight Vectors

This section explores the applicability of visualizing client ℓ_2-norms of output layer weight vectors to the CIFAR-100 dataset, which has a significantly higher number of classes than CIFAR-10 (Figure 11). Additionally, we examine whether the personalization patterns exhibited by the cwFedAvg method can be observed in other PFL algorithms such as FedAMP and FedFomo for CIFAR-10 practical settings (Figure 12). Detailed explanations are included in the figure captions.

G. Convergence Behavior Analysis

In Figure 4, we observe distinct average training loss patterns between cwFedAvg and FedAvg. We further examine the per-client convergence behaviors to analyze how different client data distributions affect the training dynamics of the two methods in Figure 13 and 14. Detailed explanations are included in the figure captions.



(a) Data distribution of clients

(b) Local models of FedAvg

(c) Local models of fine-tuned FedAvg

(d) Difference between (b) and (c)

(e) Local models of FedAMP

(f) Local models of IFCA

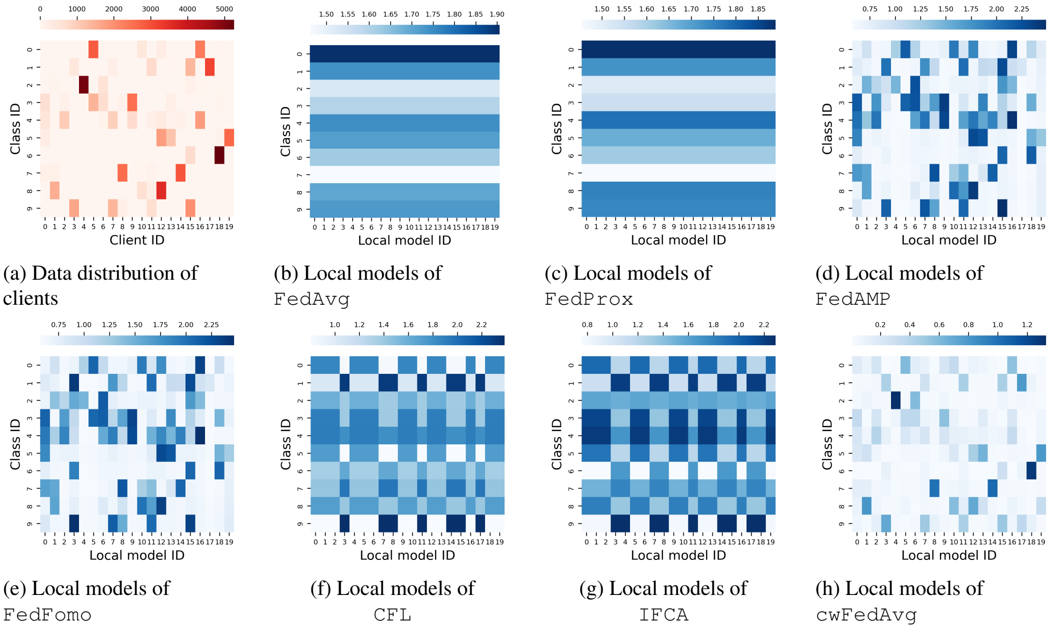
(g) Local models of cwFedAvg w/o WDR

(h) Local models of cwFedAvg w/ WDR

(i) Global models of cwFedAvg w/o WDR

(j) Global models of cwFedAvg w/ WDR

Figure 11. Heatmaps for the CIFAR-100 practical heterogeneous setting. These heatmaps confirm that the CIFAR-100 practical heterogeneous setting shows very similar patterns as the CIFAR-10 pathological heterogeneous setting. Notably, Figure 11g (cwFedAvg without WDR) closely resembles Figure 11c. In contrast, Figure 11h (cwFedAvg with WDR) exhibits a pattern similar to Figure 11a, suggesting that each model has undergone personalization tailored to its possessed classes. Additionally, we visualize ten class-specific global models of cwFedAvg in Figures 11i (without WDR) and 11j (with WDR). As designed, each global model in Figure 11j specializes in a specific single class.



(a) Data distribution of clients

(b) Local models of FedAvg

(c) Local models of FedProx

(d) Local models of FedAMP

(e) Local models of FedFomo

(f) Local models of CFL

(g) Local models of IFCA

(h) Local models of cwFedAvg

Figure 12. Heatmaps for the CIFAR-10 practical heterogeneous setting. These heatmaps confirm the observations from the CIFAR-10 pathological heterogeneous setting in Figure 3. Notably, PFL methods such as FedAMP and FedFomo exhibit patterns similar to the data distribution, albeit with less pronounced similarity compared to cwFedAvg. Interestingly, clustering-based PFL methods, such as CFL and IFCA, exhibit distinct patterns, with two clusters evident in the heatmaps. Among the various FL and PFL approaches, cwFedAvg demonstrates the most similar pattern with the true data distribution, suggesting its superior capability in personalizing clients.

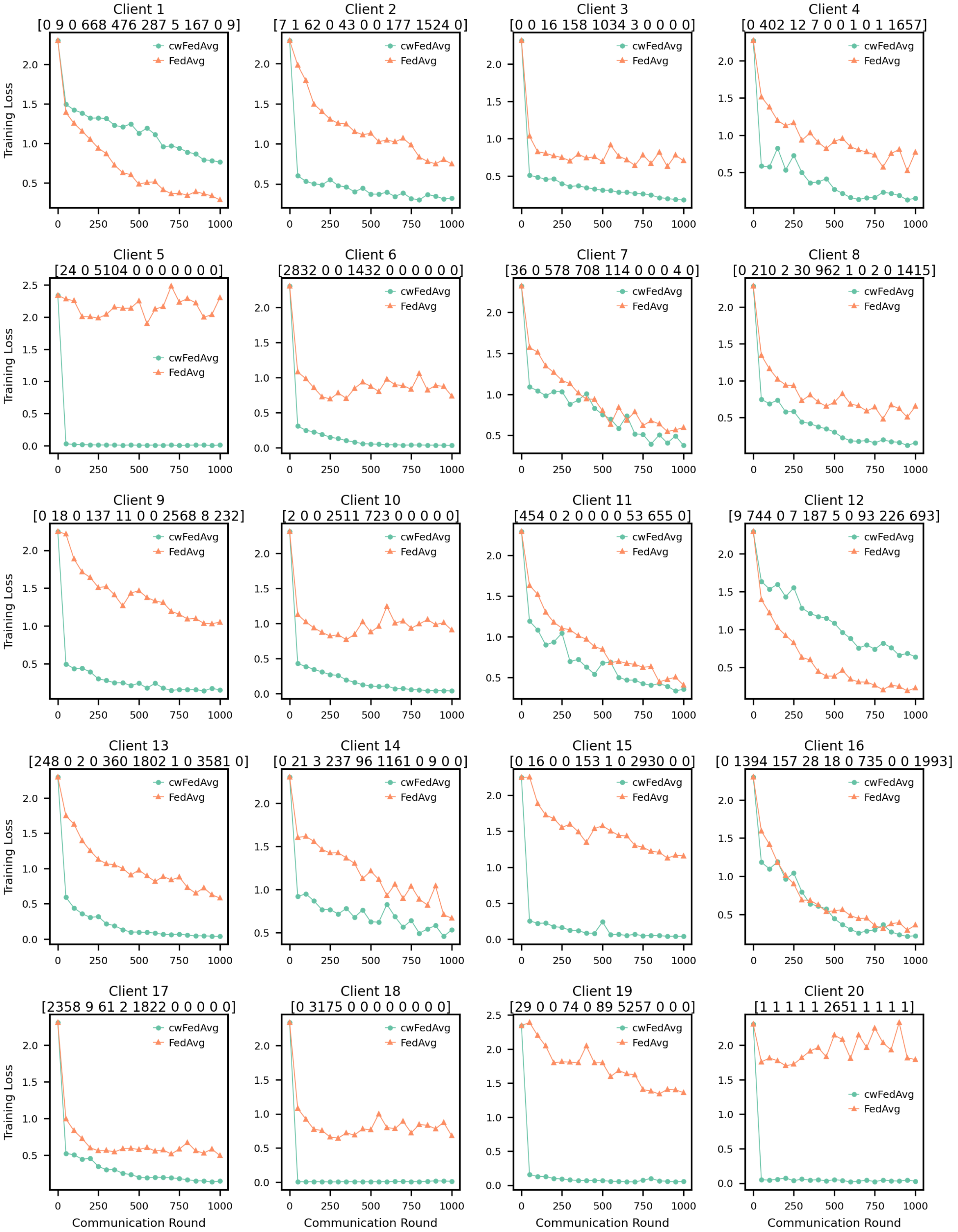


Figure 13. Comparison of Per-Client Convergence Behaviors for CIFAR-10 in Practical Settings ($\alpha = 0.1$). Figures clearly show that cwFedAvg converges significantly faster than FedAvg for highly imbalanced distributions, where the number of samples per class is shown below each client ID in the line plots. This superior convergence of cwFedAvg is observed in clients 5, 6, 9, 10, 13, 15, 18, 19, and 20. Conversely, FedAvg demonstrates faster convergence in clients 1 and 12, where the data distribution is less imbalanced.

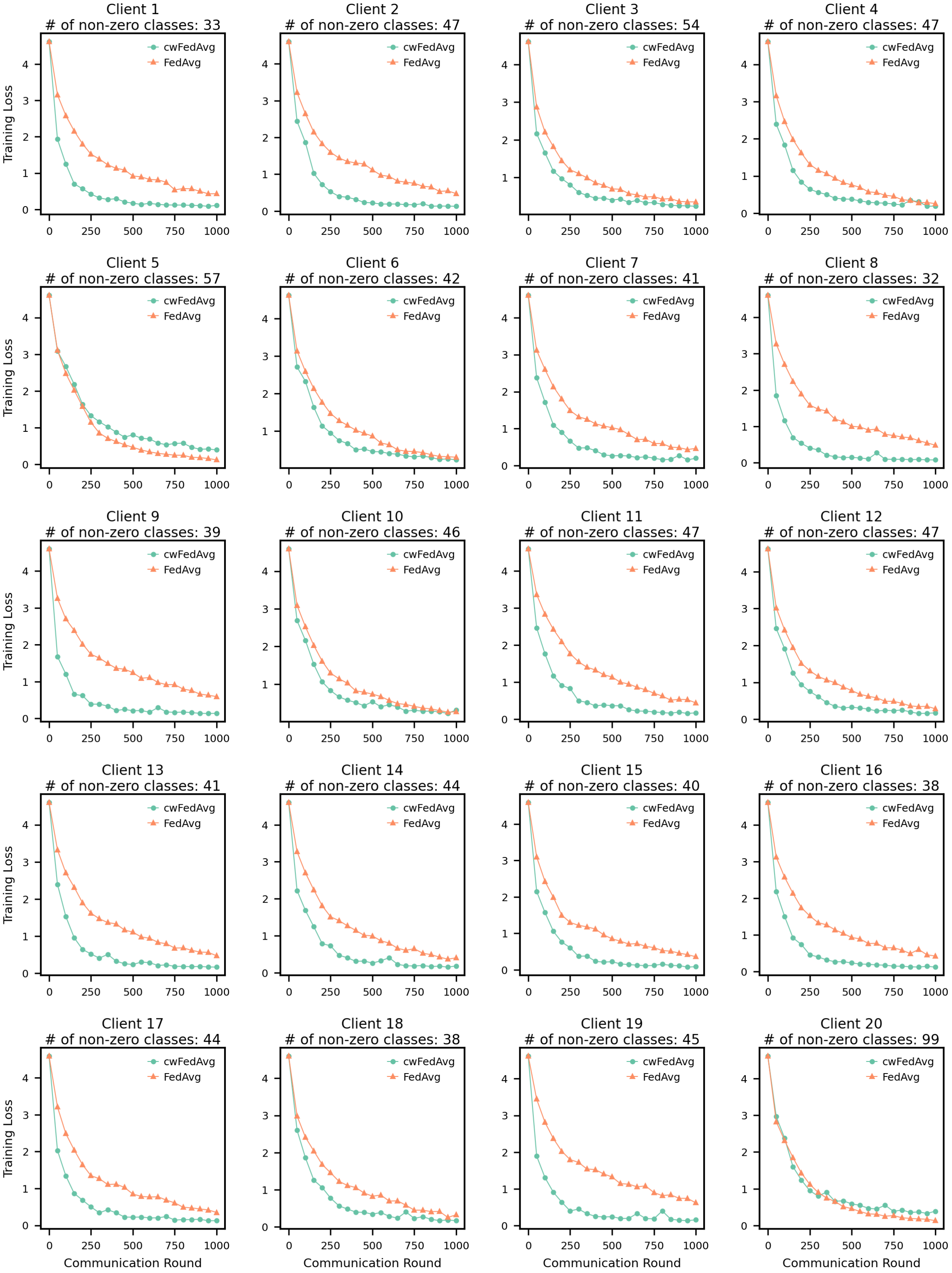


Figure 14. Comparison of Per-Client Convergence Behaviors for CIFAR-100 in Practical Settings ($\alpha = 0.1$). Figures reveal convergence characteristics that align with the findings in Figure 13. The line plots, which display the number of non-zero classes under each client ID, demonstrate that `cwFedAvg` achieves faster convergence than `FedAvg` in highly imbalanced scenarios. Although this advantage persists across most clients, `FedAvg` shows superior convergence rates for clients 5 and 20, where data is more evenly distributed.