A. Proof of Proposition 1

Definition 1 (Sample-wise deviation bound) *Let $\mathbf{x} \in \mathcal{O}_r$ denote a training example belonging to class r. The sample-wise deviation bound is given by*

$$D(\mathbf{x}) = \frac{\left(1 - P_r^{(r)}\right) \Phi_r |\mathcal{O}_r| S_r(\mathbf{x})}{\sum\limits_{j \neq r} P_r^{(j)} \Phi_j |\mathcal{O}_j| S_j(\mathbf{x})}, \tag{6}$$

where $P_z^{(y)} = \frac{1}{|\mathcal{O}_y|} \sum_{i \in \mathcal{O}_y} p_z(\mathbf{x}_i)$ is the average prediction score of the samples in a class y to a class z, $\Phi_y = \frac{1}{|\mathcal{O}_y|} \sum_{i \in \mathcal{O}_y} \|\phi(\mathbf{x}_i)\|_2$ is the average feature norm of the examples in class y, and $S_y(\mathbf{x}) = \frac{1}{|\mathcal{O}_y|} \sum_{i \in \mathcal{O}_y} \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle$ denotes the average feature similarity to a sample $\mathbf{x}$.

Proposition 1 *When $D(\mathbf{x}) \ll 1$, the local updates $\{\Delta\psi_y\}_{y \in \mathcal{Y}}$ are prone to deviate from the expected direction,* i.e., *$\Delta\psi_r \phi(\mathbf{x}) < 0$ and $\Delta\psi_j \phi(\mathbf{x}) > 0$.*

To minimize the classification error for $\mathbf{x} \in \mathcal{O}_r$, we expect $\Delta\psi_r \phi(\mathbf{x}) > 0$ and $\Delta\psi_j \phi(\mathbf{x}) < 0$ for all $j \neq r$, increasing the probability $p_r(\mathbf{x}) = \frac{\exp(\psi_r \phi(\mathbf{x}))}{\sum_{k \in \mathcal{Y}} \exp(\psi_k \phi(\mathbf{x}))}$. Following [42], we derive the update process for ψ by

$$\begin{aligned} \Delta\psi_r &= \eta \sum_{\mathbf{x}_i \in \mathcal{O}_r} \left(1 - p_r(\mathbf{x}_i)\right) \phi(\mathbf{x}_i) - \eta \sum_{j \neq r} \sum_{\mathbf{x}_i \in \mathcal{O}_j} p_r(\mathbf{x}_i) \phi(\mathbf{x}_i) \\ &\approx \eta \left(1 - P_r^{(r)}\right) \sum_{\mathbf{x}_i \in \mathcal{O}_r} \phi(\mathbf{x}_i) - \eta \sum_{j \neq r} P_r^{(j)} \sum_{\mathbf{x}_i \in \mathcal{O}_j} \phi(\mathbf{x}_i), \end{aligned} \tag{7}$$

where η is a learning rate. Then, $\Delta\psi_r \phi(\mathbf{x})$ can be formulated as

$$\begin{aligned} &\Delta\psi_r \phi(\mathbf{x}) = \eta \left(1 - P_r^{(r)}\right) \sum_{\mathbf{x}_i \in \mathcal{O}_r} \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) - \eta \sum_{j \neq r} P_r^{(j)} \sum_{\mathbf{x}_i \in \mathcal{O}_j} \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) \\ &= \eta \left(1 - P_r^{(r)}\right) \frac{\|\phi(\mathbf{x})\|_2}{|\mathcal{O}_r|} \sum_{\mathbf{x}_i \in \mathcal{O}_r} \|\phi(\mathbf{x}_i)\|_2 \sum_{\mathbf{x}_j \in \mathcal{O}_r} \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}) \rangle - \eta \sum_{j \neq r} P_r^{(j)} \frac{\|\phi(\mathbf{x})\|_2}{|\mathcal{O}_j|} \sum_{\mathbf{x}_i \in \mathcal{O}_j} \|\phi(\mathbf{x}_i)\|_2 \sum_{\mathbf{x}_j \in \mathcal{O}_j} \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}) \rangle \\ &= \eta \left(1 - P_r^{(r)}\right) \|\phi(\mathbf{x})\|_2 \, \Phi_r |\mathcal{O}_r| S_r(\mathbf{x}) - \eta \sum_{j \neq r} P_r^{(j)} \|\phi(\mathbf{x})\|_2 \, \Phi_j |\mathcal{O}_j| S_j(\mathbf{x}) \\ &= \eta \|\phi(\mathbf{x})\|_2 \left(\sum_{j \neq r} P_r^{(j)} \Phi_j |\mathcal{O}_j| S_j(\mathbf{x}) \right) \Bigg(\underbrace{\frac{\left(1 - P_r^{(r)}\right) \Phi_r |\mathcal{O}_r| S_r(\mathbf{x})}{\sum\limits_{j \neq r} P_r^{(j)} \Phi_j |\mathcal{O}_j| S_j(\mathbf{x})}}_{D(\mathbf{x})} - 1 \Bigg), \end{aligned} \tag{8}$$

where the deviation bound $D(\mathbf{x})$ in Definition 1 is derived. For the second equality, we assume that the cosine similarity of different $\phi(\mathbf{x})$ is independent with the L_2-norm of $\phi(\mathbf{x})$. In this equation, $\Delta\psi_r \phi(\mathbf{x})$ becomes negative when $D(\mathbf{x}) \ll 1$,[3] which suggests that the local updates are more likely to deviate from the expected direction with a lower value of $D(x)$.

Similarly, $\Delta\psi_j \phi(\mathbf{x})$ is described as

$$\Delta\psi_j \phi(\mathbf{x}) = \eta \|\phi(\mathbf{x})\|_2 \left(\Phi_j |\mathcal{O}_j| S_j(\mathbf{x}) - \sum_{k \in \mathcal{Y}} P_j^{(k)} \Phi_k |\mathcal{O}_k| S_k(\mathbf{x}) \right). \tag{9}$$

By taking the average of Eq. (9) over all classes excluding the class r, we get

[3] Due to the common practice of employing activation functions like ReLU, the feature output $\phi(\cdot)$ is always non-negative, and consequently, the average feature similarity $S_y(\cdot)$ is also non-negative for any $y \in \mathcal{Y}$. This indicates that the sign of $\Delta\psi_r \phi(\mathbf{x})$ is solely affected by $D(\mathbf{x})$.

$$
\begin{aligned}
\frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r}\Delta\psi_j\phi(\mathbf{x}) &= \frac{\eta\left\|\phi(\mathbf{x})\right\|_2}{|\mathcal{Y}|-1}\left(\sum_{j\neq r}\Phi_j|\mathcal{O}_j|S_j(\mathbf{x}) - \sum_{j\neq r}\sum_{k\in\mathcal{Y}}P_j^{(k)}\Phi_k|\mathcal{O}_k|S_k(\mathbf{x})\right) \\
&= \frac{\eta\left\|\phi(\mathbf{x})\right\|_2}{|\mathcal{Y}|-1}\left(\sum_{j\neq r}\Phi_j|\mathcal{O}_j|S_j(\mathbf{x}) - \sum_{k\in\mathcal{Y}}\sum_{j\neq r}P_j^{(k)}\Phi_k|\mathcal{O}_k|S_k(\mathbf{x})\right) \\
&= \frac{\eta\left\|\phi(\mathbf{x})\right\|_2}{|\mathcal{Y}|-1}\left(\sum_{k\neq r}\Phi_j|\mathcal{O}_k|S_k(\mathbf{x}) - \sum_{k\in\mathcal{Y}}(1-P_r^{(k)})\Phi_k|\mathcal{O}_k|S_k(\mathbf{x})\right) \\
&= \frac{-\eta\left\|\phi(\mathbf{x})\right\|_2}{|\mathcal{Y}|-1}\left(\Phi_r|\mathcal{O}_r|S_r(\mathbf{x}) - \sum_{k\in\mathcal{Y}}P_r^{(k)}\Phi_k|\mathcal{O}_k|S_k(\mathbf{x})\right) \\
&= \frac{-\eta\left\|\phi(\mathbf{x})\right\|_2}{|\mathcal{Y}|-1}\left(\sum_{j\neq r}P_r^{(j)}\Phi_j|\mathcal{O}_j|S_j(\mathbf{x})\right)\underbrace{\left(\frac{\left(1-P_r^{(r)}\right)\Phi_r|\mathcal{O}_r|S_r(\mathbf{x})}{\sum_{j\neq r}P_r^{(j)}\Phi_j|\mathcal{O}_j|S_j(\mathbf{x})} - 1\right)}_{D(\mathbf{x})},
\end{aligned}
\tag{10}
$$

where the same $D(\mathbf{x})$ is derived, suggesting that there exists $j \in \mathcal{Y}\setminus r$ for which $\Delta\psi_j\phi(\mathbf{x})$ becomes positive if $D(\mathbf{x}) \ll 1$. Both Eqs. (8) and (10) present that lower values of $D(\mathbf{x})$ are likely to lead to gradient deviations. □

B. Mitigating Local Gradient Deviations via SCL

By Proposition 1, our objective is improving $D(\mathbf{x})$ to prevent local gradient deviations. Assuming that $\frac{S_r(\mathbf{x})}{\sum_{j\neq r}S_j(\mathbf{x})} \geq \frac{1}{|\mathcal{Y}|-1}$, we derive the lower bound of $D(\mathbf{x})$ as

$$
\begin{aligned}
D(\mathbf{x}) &= \frac{\left(1-P_r^{(r)}\right)\Phi_r|\mathcal{O}_r|S_r(\mathbf{x})}{\sum_{k\neq r}P_r^{(k)}\Phi_k|\mathcal{O}_k|S_k(\mathbf{x})} \\
&= \frac{S_r(\mathbf{x})}{\sum_{k\neq r}\min_{j\neq r}\left\{\frac{P_r^{(k)}\Phi_k|\mathcal{O}_k|}{P_r^{(j)}\Phi_j|\mathcal{O}_j|}\right\}S_k(\mathbf{x})}(1-P_r^{(r)})\Phi_r|\mathcal{O}_r|\min_{j\neq r}\left\{\frac{1}{P_r^{(j)}\Phi_j|\mathcal{O}_j|}\right\} \\
&\geq \frac{S_r(\mathbf{x})}{\sum_{j\neq r}S_j(\mathbf{x})}(1-P_r^{(r)})\Phi_r|\mathcal{O}_r|\min_{j\neq r}\left\{\frac{1}{P_r^{(j)}\Phi_j|\mathcal{O}_j|}\right\} \qquad (11) \\
&\geq \frac{\left(1-P_r^{(r)}\right)\Phi_r|\mathcal{O}_r|}{|\mathcal{Y}|-1}\min_{j\neq r}\left\{\frac{1}{P_r^{(j)}\Phi_j|\mathcal{O}_j|}\right\}, \qquad (12)
\end{aligned}
$$

which suggests that encouraging each sample to satisfy $\frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r}S_j(\mathbf{x}) - S_r(\mathbf{x}) \leq 0$ increases the difficulty of encountering $D(\mathbf{x}) \ll 1$, thereby alleviating local gradient deviations. Thus, we formulate the surrogate objective to minimize

$$
\max\left(0, \frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r}S_j(\mathbf{x}) - S_r(\mathbf{x})\right). \tag{13}
$$

Using $\max\{a_1, \ldots, a_n\} \leq LogSumExp(a_1, \ldots, a_n)$, the upper bound of the objective is

$$
\begin{aligned}
&\max\left(0, \frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r} S_j(\mathbf{x}) - S_r(\mathbf{x})\right) \\
&\leq \log\left(\exp(0) + \exp\left(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}S_j(\mathbf{x}) - S_r(\mathbf{x})\right)\right) \\
&= \log\left(\exp(-S_r(\mathbf{x}))\left(\exp(S_r(\mathbf{x})) + \exp\left(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}S_j(\mathbf{x})\right)\right)\right) \\
&= \log\left(\exp(-S_r(\mathbf{x}))\right) + \log\left(\exp(S_r(\mathbf{x})) + \exp\left(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}S_j(\mathbf{x})\right)\right) \\
&= -\log\left(\frac{\exp(S_r(\mathbf{x}))}{\exp(S_r(\mathbf{x})) + \exp\left(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}S_j(\mathbf{x})\right)}\right) \\
&= -\log\left(\frac{\exp(\frac{1}{|\mathcal{O}_r|}\sum_{\mathbf{x}_i\in\mathcal{O}_r}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\exp(\frac{1}{|\mathcal{O}_r|}\sum_{\mathbf{x}_i\in\mathcal{O}_r}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle) + \exp(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}\frac{1}{|\mathcal{O}_j|}\sum_{\mathbf{x}_i\in O_j}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}\right) \\
&\leq -\log\left(\frac{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle) + \exp(\sum_{j\neq r}\frac{1}{|\mathcal{Y}|-1}\frac{1}{|\mathcal{O}_j|}\sum_{\mathbf{x}_i\in O_j}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}\right) \\
&\leq -\log\left(\frac{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle) + \frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r}\exp(\frac{1}{|\mathcal{O}_j|}\sum_{\mathbf{x}_i\in\mathcal{O}_j}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}\right) \qquad \text{(I*)} \\
&\leq -\log\left(\frac{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\exp(\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle) + \frac{1}{|\mathcal{Y}|-1}\sum_{j\neq r}\frac{1}{|\mathcal{O}_j|}\sum_{\mathbf{x}_i\in\mathcal{O}_j}\exp(\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}\right) \qquad \text{(II*)} \\
&\leq -\log\left(\frac{\exp(\frac{1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\sum_{\mathbf{x}_i\neq\mathbf{x}}\exp(\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}\right) \\
&= \frac{-1}{|\mathcal{O}_r|-1}\sum_{\mathbf{x}_i\in\mathcal{O}_r\setminus\mathbf{x}}\log\left(\frac{\exp(\langle\phi(\mathbf{x}),\phi(\mathbf{x}_i)\rangle)}{\sum_{\mathbf{x}_k\neq\mathbf{x}}\exp(\langle\phi(\mathbf{x}),\phi(\mathbf{x}_k)\rangle)}\right), \qquad (14)
\end{aligned}
$$

where (I*) and (II*) come from Jensen's inequality. □

C. Additional Experiments

Quantity-based data heterogeneity configurations Beside distribution-based data heterogeneity, we additionally employ quantity-based heterogeneity configurations for comprehensive evaluation. Let assume M training samples are distributed among N clients. We initially organize the data by class labels and split it into $\gamma \cdot N$ groups, with each group having $\frac{M}{\gamma \cdot N}$ samples. Note that there is no overlap in the samples held by different clients in these settings. Our framework consistently exhibits superior performance as evidenced in Table A, which verifies the robustness of our framework across diverse data heterogeneity scenarios.

Integration into server-side optimization approaches To supplement Table 5 in the main paper, we evaluate other recent client-side approaches [33, 42] combined with various server-side algorithms [9, 14, 27] for additional comparisons. As shown in Table B, our framework consistently outperforms FedLC and FedDecorr on top of existing server-side frameworks.

Other backbone networks We evaluate FedRCL using different backbone architectures, including VGG-9 [34], MobileNet-V2 [31], ShuffleNet [43], and SqueezeNet [11] on CIFAR-100, where we set β to 2 for MobileNet and 1 for others. According to Table C, FedRCL outperforms other algorithms by large margins regardless of backbone architectures, which shows the generality of our approach.

Table A. Results from quantity-based data heterogeneity configurations over 100 distributed clients on the three benchmarks.

	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	$\gamma = 2$		$\gamma = 5$		$\gamma = 20$		$\gamma = 50$		$\gamma = 20$		$\gamma = 50$	
Method	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R
FedAvg [22]	37.22	52.88	71.57	82.04	37.94	44.39	44.31	50.02	23.59	28.30	30.32	32.83
FedLC [42]	28.24	35.69	77.06	83.65	41.35	46.62	44.11	48.65	**27.90**	29.21	33.24	34.92
FedDecorr [33]	42.93	60.63	74.49	82.15	39.63	46.40	44.62	50.30	22.74	27.20	29.92	32.62
FedRCL (ours)	**55.01**	**71.66**	**81.19**	**87.66**	**51.09**	**59.78**	**58.05**	**63.50**	26.53	**33.43**	**34.18**	**41.49**

Table B. Integration of client-side approaches into various server-side approaches under non-*i.i.d.* setting ($\alpha = 0.3$).

	CIFAR-10		CIFAR-100		Tiny-ImageNet	
Method	500R	1000R	500R	1000R	500R	1000R
FedAvgM [10]	80.56	85.48	46.98	53.29	36.32	38.51
FedAvgM + FedLC	82.03	86.41	46.96	52.91	37.76	40.50
FedAvgM + FedDecorr	80.57	85.51	46.31	53.11	34.66	36.95
FedAvgM + **FedRCL (ours)**	**84.62**	**88.51**	**60.55**	**64.61**	**43.11**	**47.23**
FedADAM [27]	75.91	81.82	47.99	52.81	36.33	39.74
FedADAM + FedLC	77.96	82.11	49.76	53.15	**39.04**	42.12
FedADAM + FedDecorr	76.44	82.21	48.62	53.48	35.92	39.38
FedADAM + **FedRCL (ours)**	**80.71**	**85.69**	**52.86**	**57.84**	38.34	**42.27**
FedACG [14]	85.13	89.10	55.79	62.51	42.26	46.31
FedACG + FedLC	85.89	89.61	57.18	62.09	43.43	44.57
FedACG + FedDecorr	85.20	89.48	57.95	63.02	43.09	44.52
FedACG + **FedRCL (ours)**	**86.43**	**89.67**	**62.82**	**66.38**	**45.97**	**47.97**

Table C. Experimental results with different backbone architecture on the CIFAR-100 dataset under non-*i.i.d.* setting ($\alpha = 0.3$).

	SqueezeNet	ShuffleNet	VGG-9	MobileNet-V2
FedAvg [22]	39.62	35.37	45.60	43.57
+ FitNet [29]	37.78	36.18	45.35	43.89
FedProx [21]	38.86	35.37	45.32	43.09
MOON [19]	24.16	34.17	52.13	34.05
FedMLB [15]	41.95	41.61	54.36	47.09
FedLC [42]	42.35	37.79	48.46	45.51
FedNTD [18]	40.33	40.14	50.78	44.85
FedProc [23]	31.45	35.23	43.14	23.60
FedDecorr [33]	40.23	38.77	47.32	47.31
FedRCL (ours)	**49.34**	**44.50**	**55.53**	**51.32**

Larger number of local epochs To validate the effectiveness in conditions of more severe local deviations, we evaluate our framework by increasing the number of local epochs to $E = 10$. Table D presents consistent performance enhancements of FedRCL in the presence of more significant local deviations.

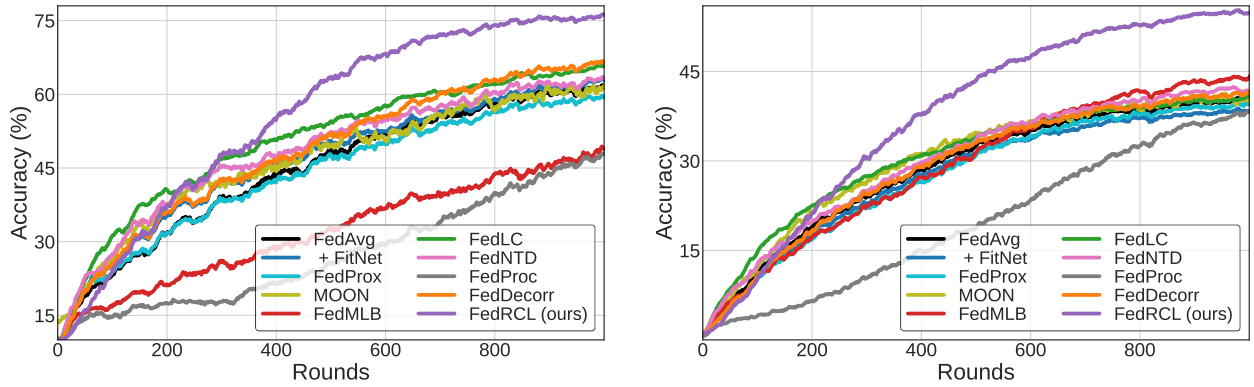
D. Qualitative Results

Convergence plot Figure A visualizes the convergence curves of FedRCL and the compared algorithms on CIFAR-10 and CIFAR-100 under non-*i.i.d.* setting ($\alpha = 0.05$), where our framework consistently outperforms all other existing federated learning techniques by huge margins throughout most of the learning process.

Sensitivity on the weight of divergence penalty We examine the robustness of our framework by varying the divergence penalty weight $\beta \in \{0, 0.1, 0.2, 0.5, 1, 2, 5\}$ on the CIFAR-100 in non-*i.i.d.* settings. Figure B presents consistent performance enhancements over a wide range of β, which demonstrates its stability.

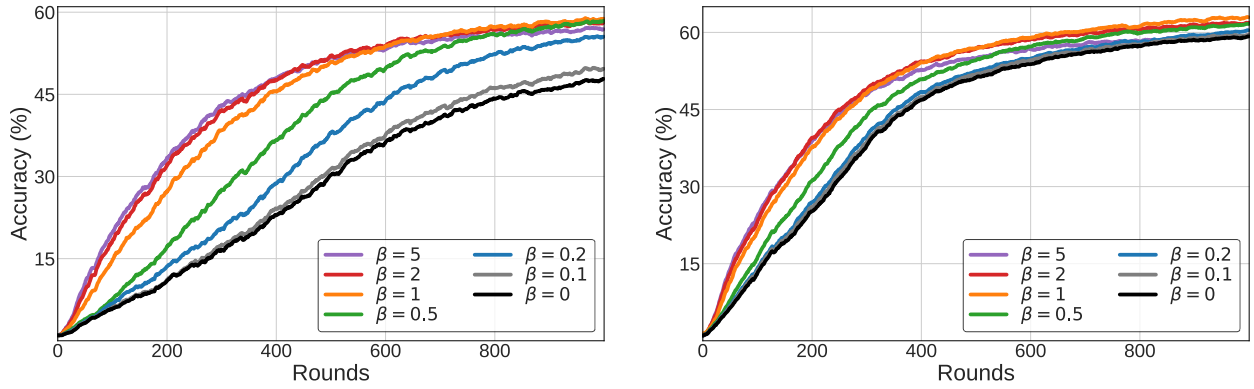
Table D. Experimental results with an increased number of local epochs ($E = 10$) under non-*i.i.d.* setting ($\alpha = 0.05$).

	CIFAR-10				CIFAR-100				Tiny-ImageNet			
	$\alpha = 0.05$		$\alpha = 0.3$		$\alpha = 0.05$		$\alpha = 0.3$		$\alpha = 0.05$		$\alpha = 0.3$	
	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R	500R	1000R
Baseline	56.80	68.52	77.79	83.78	34.64	42.35	41.47	47.49	22.38	23.65	32.49	34.58
FedLC [42]	60.81	69.59	79.58	84.71	36.83	43.99	42.7	48.04	25.73	27.51	33.38	35.30
FedDecorr [33]	58.34	68.64	80.55	84.91	34.91	41.84	42.73	49.25	21.48	22.54	30.65	33.06
FedRCL (ours)	**74.02**	**78.97**	**86.58**	**89.40**	**49.64**	**55.91**	**60.58**	**64.73**	**31.01**	**37.70**	**44.74**	**48.51**



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

Figure A. Convergence curve of FedRCL, along with other compared methods, on the CIFAR-10 and CIFAR-100 with non-*i.i.d.* setting ($\alpha = 0.05$). Accuracy at each round is based on the exponential moving average result with parameter 0.9.



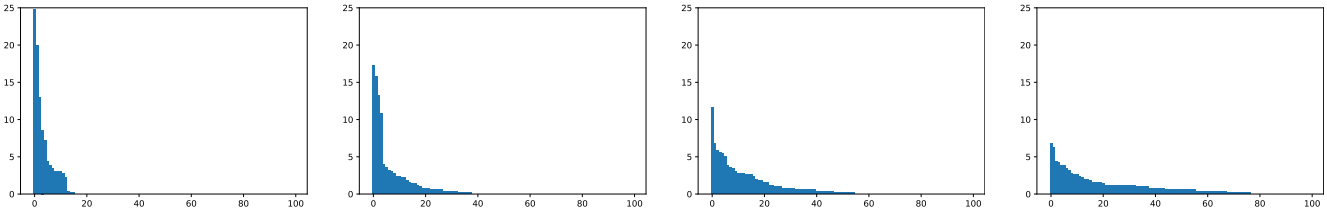
(a) $\alpha = 0.1$ (b) $\alpha = 0.3$

Figure B. Ablative results by varying the weight of the divergence penalty (β) on the CIFAR-100 dataset with $\alpha \in \{0.1, 0.3\}$, which exhibits stability across a wide range.

E. Experimental Detail

Hyperparameter selection To reproduce the compared approaches, we primarily follow the settings from their original papers, adjusting the parameters only when it leads to improved performance. In client-side federated learning approaches, we use 0.001 in FedProx, 0.3 in FedNTD, and 0.01 in FedDecorr, for β. We set λ to 0.001 in FitNet, while λ_1 and λ_2 are both set to 1 in FedMLB. μ in MOON and τ in FedLC are both set to 1. We adopt λ of 0.7, β of 1, and τ of 0.05 in FedRCL. For server-side algorithms, β in FedAvgM is set to 0.4 while β_1, β_2, and τ in FedADAM are set to 0.9, 0.99, and 0.001, respectively. We use λ of 0.85 and β of 0.001 in FedACG.

Visualization of local data distribution We visualize the local data distribution at each client on the CIFAR-100 under diverse heterogeneity configurations in Figure C, where the Dirichlet parameter α is varied by $\{0.05, 0.1, 0.3, 0.6\}$. Lower values indicate more skewed distributions.



(a) $\alpha = 0.05$ (b) $\alpha = 0.1$ (c) $\alpha = 0.3$ (d) $\alpha = 0.6$

Figure C. Label distributions at each local client under various heterogeneity configurations with $\alpha \in \{0.05, 0.1, 0.3, 0.6\}$ on the CIFAR-100. y-axis represents the ratio of data samples in each class to the total dataset, while x-axis is sorted based on the number of samples.