Additional details and results are provided in the appendices, including:

- Appendix A: Mathematical details of various representation entanglement mechanisms.
- Appendix B: Detailed experimental setup.
- Appendix C: Supplementary experimental results.

A. Mathematical Details of Various Representation Entanglement Mechanisms

We now introduce the mathematical details of different RE mechanisms. The general form of RE, calculated from a single client's representation set $\mathcal{R} = \{(\mathbf{r}_i, \mathbf{y}_i)\}_{i=1}^{n}$, is formulated by

$$\widetilde{\mathbf{r}} = \sum_{i=1}^{n} w_i \mathbf{r}_i, \widetilde{\mathbf{y}} = \sum_{i=1}^{n} w_i \mathbf{y}_i, \tag{4}$$

where $w_i \in [0, 1]$ is the weight of $\mathbf{r}_i$, which is determined by different RE mechanisms as follows:

- **Random Select Representation** (RSR) randomly selects one representation from each client per global communication round. Thus, w_i is formulated as

$$w_i = \begin{cases} 1, & \text{if } \mathbf{r}_i \text{ is selected} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

- **Vanilla Average Representation** (VAR) averages all representations per client into a single representation, with equal weight assigned to each. Hence, w_i is defined by

$$w_i = \frac{1}{n}, \forall i \in \{1, 2, \cdots, n\}. \tag{6}$$

- **Random Average Representation** (RAR) entangles representations per client into a single representation using a normalized weight vector, with elements randomly drawn from a Uniform distribution $\mathcal{U}(0, 1)$ and normalized to sum to one. Accordingly, w_i is formulated as follows:

$$w_i = \frac{u_i}{\sum_{j=1}^{n} u_j}, \text{where } u_i \sim \mathcal{U}(0, 1). \tag{7}$$

- **Random Select Prototype** (RSP) first calculates prototypes for each client and then randomly selects one prototype per client in each global communication round. Therefore, w_i is defined as

$$w_i = \begin{cases} \frac{1}{n_c}, & \text{if both the selected prototype and } \mathbf{r}_i \text{belong to category } c \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

  where n_c denotes the total number of samples belonging to category c.
- **Vanilla Average Prototype** (VAP) calculates prototypes for each client and averages them into a single representation, where each prototype contributes equally. Thus, w_i is calculated by

$$w_i = \frac{1}{C n_c}, \text{if } \mathbf{r}_i \text{belongs to category } c, \tag{9}$$

  where C is the total number of categories in the client.
- **Random Average Prototype** (RAP) calculates prototypes for each client and aggregates them into a single representation using a normalized random weight vector, where each weight $u_c \sim \mathcal{U}(0, 1)$, and the weights are normalized to sum to one. Using those weights, w_i is defined as follows:

$$w_i = \frac{u_c}{n_c \sum_{j=1}^{C} u_j}, \quad \text{if } \mathbf{r}_i \text{ belongs to category } c. \tag{10}$$

B. Detailed Experimental Setup

Table 10 provides a detailed description of the experimental setup used in this paper, covering statistical-heterogeneous settings, model training details, and model configurations. In addition, we detail the representation inversion attack setup. We assume a *semi-honest* server with full access to the representation extractor $\boldsymbol{g}(\boldsymbol{\phi}; \cdot)$, attempting to reconstruct clients' original samples

via a representation inversion attack. Following [37], given a target vector $\mathbf{r}$ (*e.g.*, representation, prototype, or entangled representation), the server optimizes an inversion network $\boldsymbol{I}(\psi;\cdot)$ with fixed noise $\mathbf{z}$ by solving $\min_{\psi} \|\boldsymbol{g}(\boldsymbol{\phi};\boldsymbol{I}(\psi;\mathbf{z})) - \mathbf{r}\|_2^2$. To ensure a fair comparison, we use a single four-layer CNN as the representation extractor and adopt the same inversion network as in [37] for all reconstruction experiments.

Table 10. *Detailed experimental setup used in this paper.*

Statistic-heterogeneous Settings	Practical setting (PRA); Pathological setting (PAT)
Model Training Details	Local batch size: 64 (TinyImageNet), 32 (CIFAR-10 & CIFAR-100) Local optimizer: SGD Local learning rate η_l: 0.06 (Model-Heterogeneity with PRA & PAT, CIFAR-10/100/TinyImageNet) 0.01 (Model-Homogeneity with PRA & PAT, CIFAR-100/TinyImageNet) 0.007 (Model-Homogeneity with PRA, CIFAR-10) 0.008 (Model-Homogeneity with PAT, CIFAR-10) Server batch size: 10; Server optimizer: SGD; Server learning rate: 0.01
Model Configurations	Local model in Model-Heterogeneity: CNN; MobileNetV2; GoogleNet; ResNet-18/34/50/101/152; ViT-B/16; ViT-B/32 Local model in Model-Homogeneity: CNN (CIFAR-10/100); ResNet-18 (TinyImageNet)

C. Supplementary Experimental Results

C.1. Model-homogeneous FL Evaluation

Model-homogeneous FL can be regarded as a special case of model-heterogeneous FL, where all clients adopt the same model architecture. In our experiments, we adopt a four-layer CNN for CIFAR-10 and CIFAR-100, and ResNet-18 for TinyImageNet across all methods. Table 11 presents the results, where FedRE achieves the best performance across all datasets. Specifically, FedRE's average accuracy is 63.21%, outperforming the second-best method, *i.e.*, LG-FedAvg, by 2.58%. Figure 5 provides convergence comparisons on TinyImageNet, where FedRE exhibits a rapid initial improvement followed by gradual stabilization, indicating stable convergence behavior throughout training. Those results suggest that FedRE remains effective in model-homogeneous FL.

Table 11. Accuracy (%) comparison on three datasets in the model-homogeneous setting. In each column, the best results are **bolded**, and the second-best results are underlined.

Method	**PRA**			**PAT**			**Average**
	CIFAR-10	**CIFAR-100**	**TinyImageNet**	**CIFAR-10**	**CIFAR-100**	**TinyImageNet**	
LG-FedAvg [21]	86.92 ± 0.25	49.82 ± 0.39	32.00 ± 0.13	90.59 ± 0.17	66.00 ± 0.27	38.43 ± 0.23	60.63
FedAvg [25]	55.21 ± 0.12	30.37 ± 0.02	13.66 ± 0.41	52.70 ± 0.11	24.89 ± 0.20	9.98 ± 0.48	31.14
FedALA [55]	55.02 ± 0.14	29.89 ± 0.22	13.63 ± 0.10	52.83 ± 0.19	24.91 ± 0.15	10.65 ± 0.15	31.16
FedGH [48]	86.02 ± 0.17	48.59 ± 0.60	28.64 ± 0.26	90.46 ± 0.22	65.14 ± 0.26	32.40 ± 0.19	58.54
FedKD [42]	86.23 ± 0.12	51.91 ± 0.28	29.47 ± 0.31	90.01 ± 0.09	67.23 ± 0.38	35.34 ± 0.33	60.03
FedAvgDBE [54]	78.10 ± 0.20	35.23 ± 0.24	16.92 ± 0.52	82.27 ± 0.45	35.21 ± 0.27	16.80 ± 0.23	44.09
FedGen [59]	55.21 ± 0.14	29.90 ± 0.17	13.76 ± 0.23	52.37 ± 0.22	24.82 ± 0.38	10.67 ± 0.54	31.12
FedProto [36]	85.63 ± 0.22	50.52 ± 0.19	28.67 ± 0.17	91.04 ± 0.16	69.28 ± 0.07	34.75 ± 0.49	59.98
FPL [11]	83.60 ± 0.03	49.10 ± 0.21	26.87 ± 0.09	90.59 ± 0.06	67.31 ± 0.03	32.95 ± 0.23	58.40
FedMRL [49]	82.55 ± 0.31	48.41 ± 0.09	26.78 ± 0.05	89.02 ± 0.23	65.97 ± 0.21	35.22 ± 0.05	57.99
FedTGP [56]	85.59 ± 0.12	47.05 ± 0.17	30.89 ± 0.21	90.49 ± 0.03	67.47 ± 0.07	40.88 ± 0.20	60.40
Local	86.33 ± 0.11	49.88 ± 0.40	31.44 ± 0.15	90.54 ± 0.12	66.57 ± 0.17	37.46 ± 0.27	60.37
FedRE	**86.99 ± 0.01**	**52.12 ± 0.04**	**36.12 ± 0.21**	**91.06 ± 0.01**	**70.52 ± 0.17**	**42.45 ± 0.17**	**63.21**

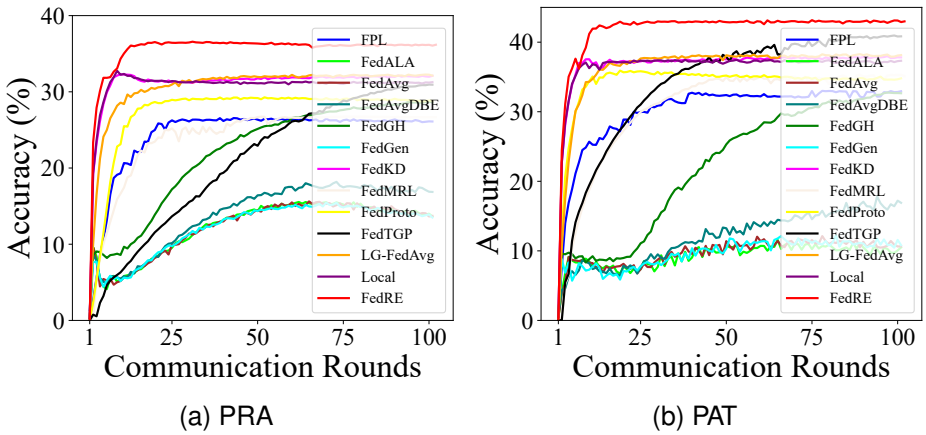


Figure 5. Accuracy (%) comparison between distinct communication rounds on the TinyImageNet dataset in the model-homogeneous FL setting in both the PRA and PAT settings.

C.2. Impact of a Small Number of Clients

We conduct experiments on CIFAR-10 and CIFAR-100 to evaluate the impact of a small number of clients ($K \in \{2, 4, 10\}$) in the model-homogeneous setting, using a four-layer CNN. Table 12 reports the results, from which we draw several observations. (1) Both FedGH and FedRE experience performance degradation as K decreases, due to reduced diversity in the prototypes or entangled representations used to train the global classifier. (2) Despite this trend, FedProto, FedGH, and FedRE consistently outperform FedAvg, demonstrating their effectiveness even with fewer clients. (3) FedRE achieves the best performance across all settings. Specifically, with only 2 clients, it reaches 41.57% on CIFAR-100, outperforming FedAvg by 8.44%, FedProto by 3.70%, and FedGH by 1.96%. Those results indicate that FedRE remains effective even with a small number of clients.

Table 12. Accuracy (%) comparison on CIFAR-10 and CIFAR-100 with different numbers of clients. In each column, the best results are **bolded**, and the second-best results are underlined.

Method	CIFAR-10			CIFAR-100		
	2	4	10	2	4	10
FedAvg [25]	64.99	62.26	55.21	33.13	30.88	30.37
FedProto [36]	77.09	<u>84.69</u>	85.63	37.87	44.85	<u>50.52</u>
FedGH [48]	<u>78.80</u>	84.37	<u>86.02</u>	<u>39.61</u>	<u>46.40</u>	48.59
FedRE	**79.11**	**85.09**	**86.99**	**41.57**	**47.86**	**52.12**

C.3. Privacy Protection Evaluation

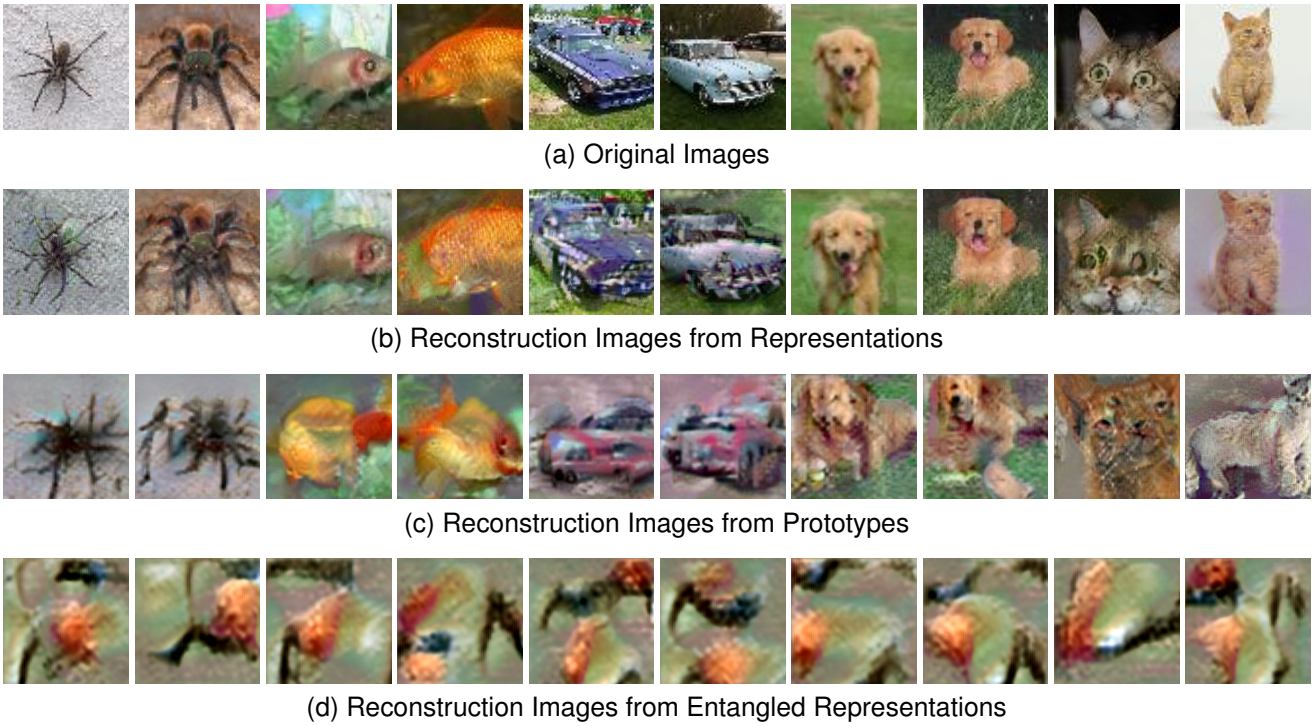
Figure 6 presents additional reconstructed results on sample images from the TinyImageNet dataset. As can be seen, images reconstructed from the entangled representations contain less discernible content or category information than those reconstructed from vanilla representations or prototypes, suggesting that FedRE offers a good level of privacy protection against representation inversion attacks.

C.4. Communication Overhead Evaluation

Table 13 lists the communication overhead results on the CIFAR-10, CIFAR-100, and TinyImageNet datasets under both model-heterogeneous (Model-hete) and model-homogeneous (Model-homo) scenarios in the PRA setting. As can be seen, FedRE generally achieves lower communication overhead than the baselines, which suggests its potential effectiveness in reducing communication overhead.

C.5. Statistical Heterogeneity Analysis

To further evaluate the effectiveness of FedRE under different levels of statistical heterogeneity, we adjust the Dirichlet distribution parameter α (*i.e.*, 0.05, 0.1, 1, 10) in the PRA setting and the client participation rate (*i.e.*, 5/25, 10/25 for 25



(a) Original Images

(b) Reconstruction Images from Representations

(c) Reconstruction Images from Prototypes

(d) Reconstruction Images from Entangled Representations

Figure 6. Comparison of privacy protection in restructuring results from representations, prototypes, and entangled representations on the TinyImageNet dataset.

Table 13. Communication overhead (# Scalars $\times 10^3$) comparison on three datasets. In each column, the best results are **bolded**, and the second-best results are underlined.

Method	CIFAR-10				CIFAR-100				TinyImageNet			
	Model-homo		Model-hete		Model-homo		Model-hete		Model-homo		Model-hete	
	Upload	Broadcast	Upload	Broadcast	Upload	Broadcast	Upload	Broadcast	Upload	Broadcast	Upload	Broadcast
LG-FedAvg	51.30	51.30	51.30	51.30	513.00	513.00	513.00	513.00	4098.00	4098.00	4098.00	4098.00
FedGH	31.23	**51.20**	31.23	**51.20**	257.02	**512.00**	257.02	**512.00**	1918.98	**4096.00**	1918.98	**4096.00**
FedKD	3374.28	3374.28	3353.68	3353.68	4234.28	4234.28	3524.67	3524.67	90503.00	90503.00	57544.97	57544.97
FedGen	8785.38	8785.38	51.30	51.30	9247.08	9247.08	513.00	513.00	239178.32	239178.32	4098.00	4098.00
FedProto	31.23	**51.20**	31.23	**51.20**	257.02	**512.00**	257.02	**512.00**	1918.98	**4096.00**	1918.98	**4096.00**
FPL	31.23	87.04	31.23	112.64	257.02	916.48	257.02	1182.72	1918.98	9768.96	1918.98	10567.68
FedMRL	8746.98	8746.98	8746.98	8746.98	8863.08	8863.08	8863.08	8863.08	56178.00	56178.00	56178.00	56178.00
FedTGP	31.23	**51.20**	31.23	**51.20**	257.02	**512.00**	257.02	**512.00**	1918.98	**4096.00**	1918.98	**4096.00**
FedAvg	8785.38	8785.38	-	-	9247.08	9247.08	-	-	239178.32	239178.32	-	-
FedALA	8785.38	8785.38	-	-	9247.08	9247.08	-	-	239178.32	239178.32	-	-
FedAvgDBE	8785.38	8785.38	-	-	9247.08	9247.08	-	-	239178.32	239178.32	-	-
FedRE	**5.12**	51.30	**5.12**	51.30	**5.12**	513.00	**5.12**	513.00	**20.48**	4098.00	**20.48**	4098.00

clients, 5/10, 10/10 for 10 clients) in the PAT setting, respectively, to control the degree of sample skewness. The resulting sample distributions are visualized in Figures 7-8. The results on the CIFAR-10 and CIFAR-100 datasets, under the model-heterogeneous setting, are shown in Figure 9. As shown, FedRE achieves competitive accuracy across different levels of statistical heterogeneity, suggesting it remains effective under varying degrees of data heterogeneity.

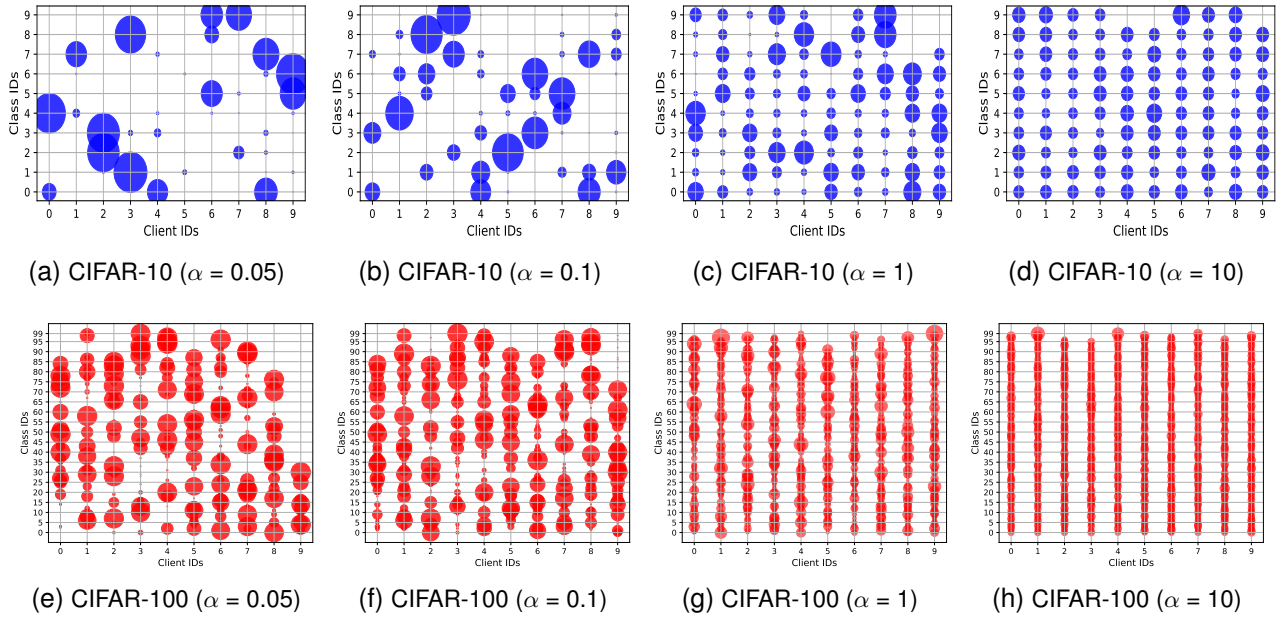


Figure 7. The sample distributions for all clients on the CIFAR-10 and CIFAR-100 datasets under the PRA settings with varying parameters α. The size of each circle indicates the number of samples.

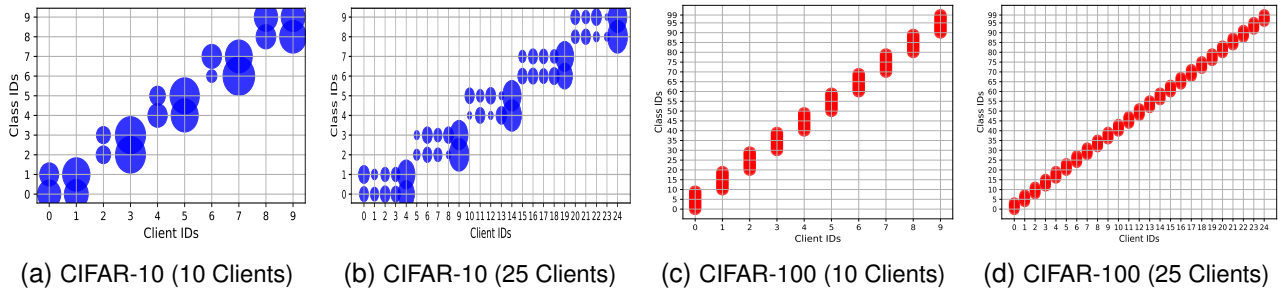


Figure 8. The sample distributions for all clients on the CIFAR-10 and CIFAR-100 datasets under the PAT settings with varying client numbers. The size of each circle indicates the number of samples.

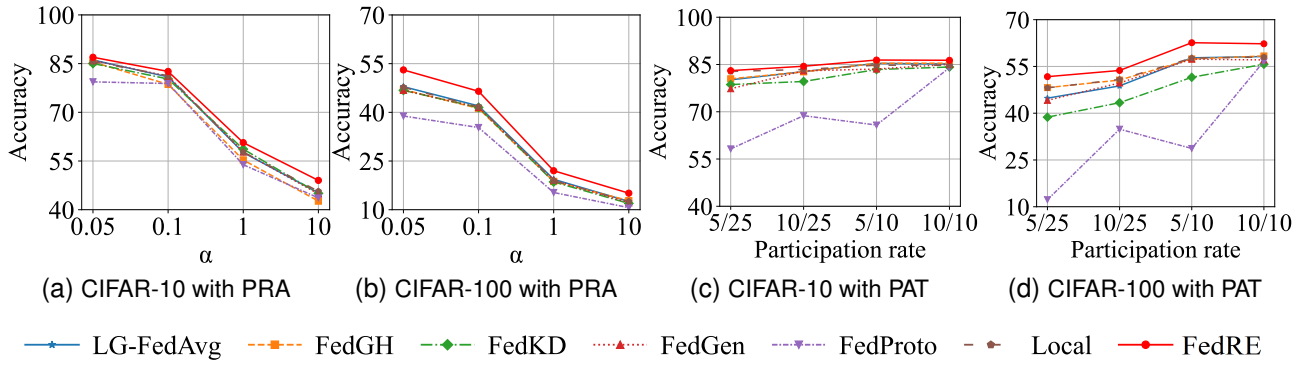


Figure 9. Accuracy (%) comparison between distinct statistic-heterogeneous scenarios on the CIFAR-10 and CIFAR-100 datasets.