# FedVG: Gradient-Guided Aggregation for Enhanced Federated Learning

## Appendix

### A.1. Relationship Between Gradient Norms and Joint Fisher in the Zero-Limit

Let the model conditional distribution be $p_\theta(y \mid x)$. For the cross-entropy loss

$$\ell(\theta; x, y) = -\log p_\theta(y \mid x), \tag{E.1}$$

the gradient is

$$\nabla_\theta \ell(\theta; x, y) = -\nabla_\theta \log p_\theta(y \mid x) = -s(x, y; \theta), \tag{E.2}$$

where $s(x, y; \theta)$ is the score function.
The (population) Fisher Information Matrix is

$$F(\theta) = \mathbb{E}[\, ss^\top \,], \tag{E.3}$$

and the Joint Fisher is defined element-wise by

$$J(\theta) = \mathbb{E}[\, s \circ s \,]. \tag{E.4}$$

Using $\|s\|_2^2 = s^\top s = \mathrm{tr}(ss^\top)$, we have

$$\mathbb{E}[\|s\|_2^2] = \mathrm{tr}(F(\theta)) = \sum_j J_j(\theta). \tag{E.5}$$

Because $\|s\|_2^2 \geq 0$, the equality

$$\mathrm{tr}(F(\theta)) = 0 \quad \Longleftrightarrow \quad \mathbb{E}[\|s\|_2^2] = 0. \tag{E.6}$$

holds if and only if $\|s(x, y; \theta)\|_2 = 0$ almost surely.
Consequently,

$$\nabla_\theta \ell(\theta; x, y) = 0 \qquad \text{and} \qquad J(\theta) = 0. \tag{E.7}$$

Thus, for cross-entropy, vanishing Joint Fisher is equivalent to vanishing gradient norms, showing that both quantities measure the same "zero-sensitivity" or "flatness" condition of the model i.e., the parameters no longer respond to perturbations in the data.

### A.2. Behaviors of Layers in an FL Setting

Fig. A.1 presents heatmaps of client-wise gradient norms across a selected subset of network layers at two training rounds. We observe that the later layers consistently exhibit larger gradient norms, reflecting their greater sensitivity to client-specific updates. In contrast, earlier layers show smaller and more uniform magnitudes. The variation across layers highlights distinct gradient behaviors within the network.

### A.3. Additional Details of Experimental Settings

This section presents additional details on the hyperparameter configurations used in our experiments.

#### A.3.1. Details of Hyperparameter Configurations

Table A.1 summarizes the hyperparameter configurations used in our federated learning experiments for each dataset. While certain parameters, such as the optimizer (SGD), learning rate (0.01), number of communication rounds (200), and local epochs (5), are consistent across datasets, others are tailored to dataset characteristics. While the number of training rounds is fixed at 200, model selection is performed to identify the best round and corresponding model based on performance. Table A.2 lists the baseline-specific hyperparameters used in our experiments.

#### A.3.2. Examples of Non-IID Client Distribution

Figure A.2 illustrates the CIFAR-10 class distributions across clients for different levels of data heterogeneity, controlled by the Dirichlet concentration parameter $\alpha$. Lower values of $\alpha$ corresponds to more skewed label distributions, resulting in higher non-IID heterogeneity across clients. Figure A.3 illustrates the COVID19 class distributions across clients for different values of $\alpha$.

### A.4. Additional Results of FedVG Performance

In this section, we provide detailed numerical results of various federated learning algorithms in tabular form to facilitate clearer comparisons.

#### A.4.1. Performance on ResNet Model Architectures

Fig. A.4 shows the performances of FedVG and baseline algorithms for Tiny-ImageNet and DermaMNIST datasets on ResNet-50 model. Table A.3 summarizes the overall performance of various federated learning methods on ResNet-based models across different dataset-model combinations and degrees of data heterogeneity.

#### A.4.2. Performance on Vision Transformer Model Architectures

Table A.4 summarizes the overall performance of various federated learning methods on ViT-based models across different dataset-model combinations and degrees of data heterogeneity.

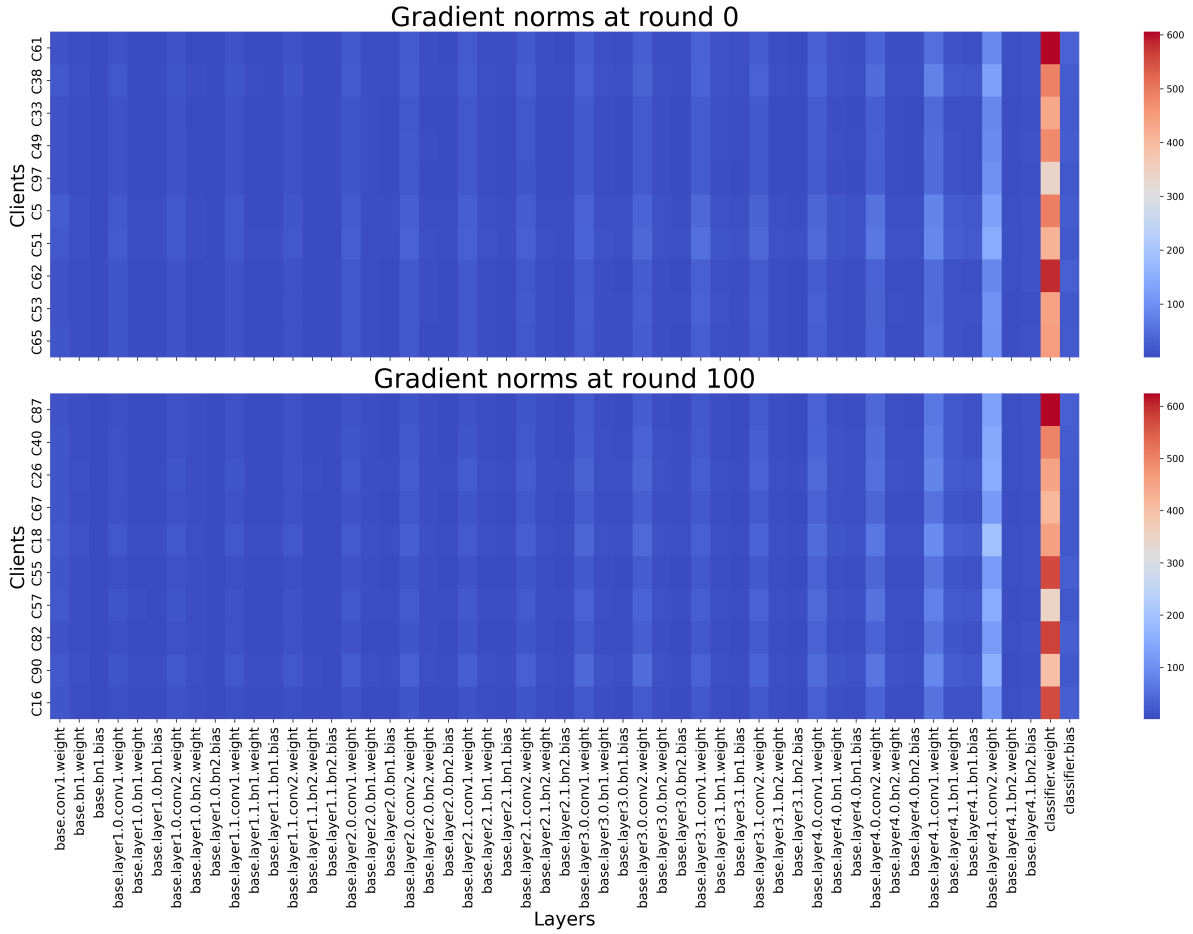


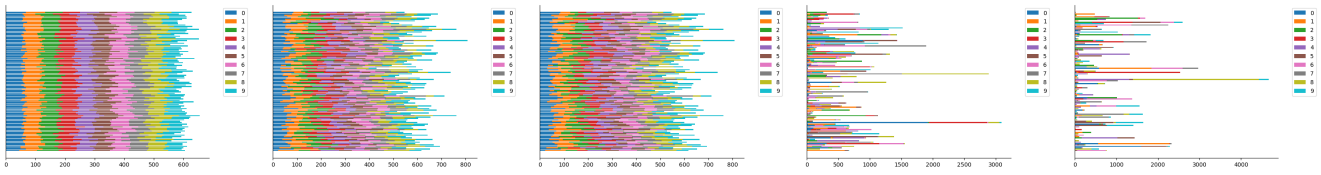

Figure A.1. Heatmaps of per-client gradient norms across network layers for two federated learning rounds. Only a subset of layers is presented for clarity.

Table A.1. Hyperparameter configurations used for different datasets in the federated learning experiments.

| | CIFAR-10 | OrganAMNIST | Tiny ImageNet | COVID19 | DermaMNIST |
|---|---|---|---|---|---|
| Number of Clients | 100 | 100 | 100 | 20 | 25 |
| Join Ratio | 0.1 | 0.1 | 0.1 | 0.25 | 0.2 |
| Input Channels | 3 | 1 | 3 | 3 | 3 |
| Image Size | 32×32 | 28×28 | 64×64 | 244×224 | 224×224 |
| Batch Size | 32 | 32 | 128 | 32 | 32 |
| Optimizer | SGD | SGD | SGD | SGD | SGD |
| Learning Rate | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| Momentum | 0 | 0 | 0.9 | 0 | 0 |
| Number of Rounds | 200 | 200 | 200 | 200 | 200 |
| Local Epochs | 5 | 5 | 5 | 5 | 5 |

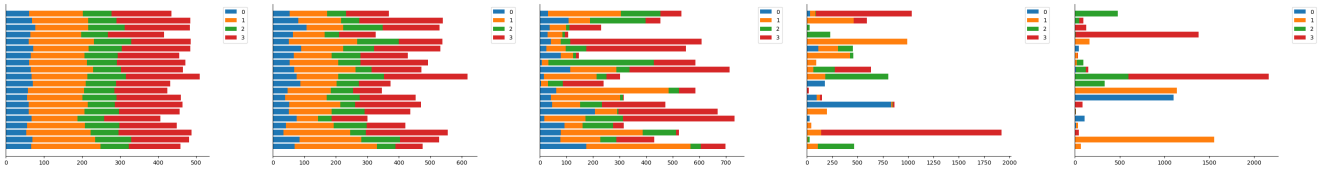
### A.4.3. FedVG Performances at High Heterogeneity ($\alpha = 0.05$)

Fig. A.5 presents a grouped bar plot comparing the performance of various federated learning methods at a high level of data heterogeneity ($\alpha = 0.05$) across seven dataset–model combinations. FedVG consistently achieves high average accuracy while maintaining relatively low variance across different tasks, demonstrating its robustness to extreme non-IID settings. Elastic aggregation also



(a) $\alpha = 100$ (b) $\alpha = 10$ (c) $\alpha = 1$ (d) $\alpha = 0.1$ (e) $\alpha = 0.05$

Figure A.2. CIFAR-10 class distributions at different heterogeneity levels, parameterized by $\alpha$ in the Dirichlet distribution. Lower $\alpha$ indicates greater heterogeneity. Each row corresponds to a client, and colors represent different classes.



(a) $\alpha = 100$ (b) $\alpha = 10$ (c) $\alpha = 1$ (d) $\alpha = 0.1$ (e) $\alpha = 0.05$

Figure A.3. COVID19 class distributions at different heterogeneity levels, parameterized by $\alpha$ in the Dirichlet distribution. Lower $\alpha$ indicates greater heterogeneity. Each row corresponds to a client, and colors represent different classes.

Table A.2. Baseline-specific hyperparameters used in our experiments.

| Baseline | Hyperparameter | Value |
|---|---|---|
| FedAvgM | server momentum | 0.9 |
| FedProx | $\mu$ | 0.01 |
| Scaffold | global lr | 1.0 |
| FedDyn | $\alpha$ | 0.1 |
| | max_grad_norm | 10 |
| Elastic | sample_ratio | 0.3 |
| | $\tau$ | 0.5 |
| | $\mu$ | 0.95 |

emerges as a competitive baseline, achieving strong results in several cases, though with slightly higher variability in certain datasets.

### A.4.4. Additional Comparison with FedAWA

We conduct additional experiments, requested during the rebuttal phase, comparing FedVG with the recent baseline FedAWA [31] across different levels of data heterogeneity. Table A.5 summarizes the results under $\alpha \in 0.05, 0.1, 1, 10, 100$. In particular, under extreme heterogeneity ($\alpha = 0.05$), FedVG achieves a higher mean accuracy (73.23) compared to FedAWA (72.61), indicating improved robustness in challenging settings.

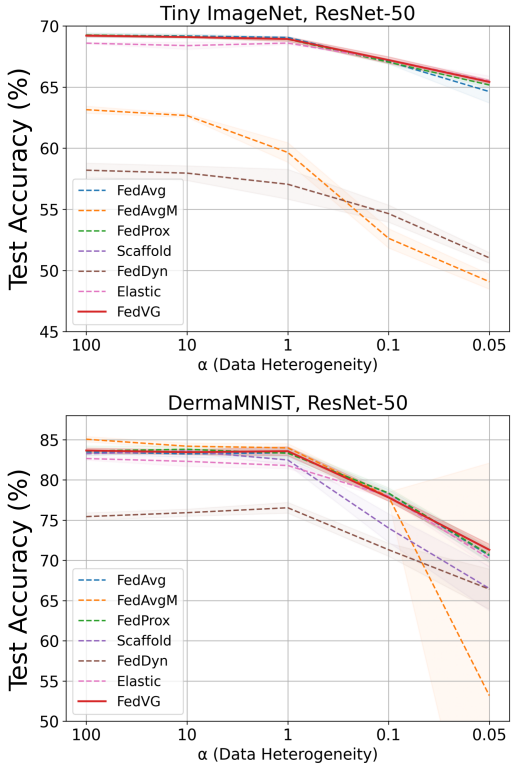



Figure A.4. FL algorithm performance for TinyImageNet and DermaMNIST datasets on the ResNet-50 model as $\alpha \to 0$. Shaded areas indicate standard deviations.

## A.5. Validation set curation using public dataset

For the experiments in Section 6, we curated the validation set using CIFAR-100 and STL-10. Tables A.7 and A.6

Table A.3. Test accuracy (%) ± standard deviation of various federated learning methods on CIFAR-10, OrganAMNIST, Tiny ImageNet, COVID19, and DermaMNIST datasets across different data heterogeneity levels (Dirichlet $\alpha$). **Bold** values indicate best accuracies and underlined values indicate second best accuracies.

| Methods | $\alpha = 100$ | $\alpha = 10$ | $\alpha = 1$ | $\alpha = 0.1$ | $\alpha = 0.05$ |
|---|---|---|---|---|---|
| *CIFAR-10 / ResNet-18* | | | | | |
| FedAvg | 80.37 ± 0.89 | 77.95 ± 0.43 | 76.26 ± 1.65 | 57.93 ± 1.82 | 48.83 ± 5.53 |
| FedAvgM | 71.09 ± 5.16 | 56.64 ± 10.75 | 51.87 ± 11.40 | 34.18 ± 4.38 | 25.87 ± 7.04 |
| FedProx | 79.92 ± 1.50 | 77.39 ± 1.32 | 76.73 ± 0.63 | 55.22 ± 5.56 | 49.19 ± 6.55 |
| Scaffold | **81.05 ± 0.16** | **80.39 ± 0.15** | **79.25 ± 0.23** | 56.15 ± 2.65 | 41.94 ± 1.91 |
| FedDyn | 66.37 ± 0.33 | 65.90 ± 0.36 | 62.37 ± 0.29 | 41.66 ± 1.22 | 39.26 ± 1.99 |
| Elastic | 79.45 ± 0.27 | 76.55 ± 1.88 | 75.69 ± 0.89 | 55.86 ± 4.26 | **54.92 ± 2.16** |
| FedVG (Ours) | 80.74 ± 0.79 | 79.65 ± 0.52 | 75.69 ± 2.27 | **61.06 ± 0.34** | 53.58 ± 2.78 |
| *OrganAMNIST / ResNet-18* | | | | | |
| FedAvg | 98.46 ± 0.16 | 98.87 ± 0.21 | 98.47 ± 0.24 | 93.50 ± 1.22 | 86.37 ± 2.49 |
| FedAvgM | 87.14 ± 4.56 | 93.43 ± 4.13 | 94.00 ± 4.76 | 78.95 ± 8.14 | 60.15 ± 7.61 |
| FedProx | 98.44 ± 0.21 | 98.68 ± 0.32 | 98.60 ± 0.20 | 93.19 ± 1.63 | 83.80 ± 1.25 |
| Scaffold | 99.22 ± 0.06 | 99.17 ± 0.06 | 98.85 ± 0.06 | 93.11 ± 0.62 | 84.32 ± 1.72 |
| FedDyn | 97.26 ± 0.17 | 97.13 ± 0.07 | 95.60 ± 0.15 | 78.61 ± 1.01 | 68.17 ± 0.62 |
| Elastic | 99.12 ± 0.11 | 98.78 ± 0.18 | 97.94 ± 0.27 | 88.46 ± 3.70 | 79.96 ± 5.20 |
| FedVG (Ours) | **99.41 ± 0.08** | **99.42 ± 0.02** | **99.12 ± 0.10** | **94.72 ± 0.68** | **87.57 ± 1.91** |
| *Tiny ImageNet / ResNet-50* | | | | | |
| FedAvg | 69.25 ± 0.19 | **69.18 ± 0.16** | **69.06 ± 0.08** | 67.08 ± 0.18 | 64.63 ± 0.93 |
| FedAvgM | 63.15 ± 0.29 | 62.67 ± 0.14 | 59.65 ± 0.80 | 52.63 ± 0.79 | 49.08 ± 0.61 |
| FedProx | **69.26 ± 0.14** | 69.13 ± 0.17 | 68.91 ± 0.16 | 67.04 ± 0.16 | 65.17 ± 0.22 |
| Scaffold | 40.53 ± 2.33 | 40.55 ± 2.53 | 37.85 ± 3.29 | 34.40 ± 5.73 | 29.23 ± 3.54 |
| FedDyn | 58.20 ± 0.58 | 57.96 ± 0.60 | 57.05 ± 1.22 | 54.65 ± 0.70 | 51.03 ± 0.43 |
| Elastic | 68.59 ± 0.05 | 68.39 ± 0.30 | 68.60 ± 0.11 | **67.23 ± 0.04** | **65.50 ± 0.40** |
| FedVG (Ours) | 69.21 ± 0.10 | 69.09 ± 0.05 | 68.93 ± 0.18 | 67.21 ± 0.27 | 65.42 ± 0.18 |
| *COVID19 / ResNet-50* | | | | | |
| FedAvg | 88.23 ± 0.23 | 87.88 ± 0.38 | 87.91 ± 0.37 | 84.34 ± 0.67 | 64.92 ± 5.45 |
| FedAvgM | **90.16 ± 0.23** | **90.23 ± 0.54** | **90.00 ± 0.33** | **85.54 ± 0.87** | 60.07 ± 9.31 |
| FedProx | 87.93 ± 0.55 | 87.91 ± 0.37 | 87.53 ± 0.47 | 84.06 ± 1.06 | 64.58 ± 5.34 |
| Scaffold | 88.70 ± 0.30 | 88.67 ± 0.55 | 88.36 ± 0.26 | 84.47 ± 1.06 | 70.05 ± 5.29 |
| FedDyn | 84.48 ± 0.48 | 84.51 ± 0.33 | 84.28 ± 0.75 | 78.49 ± 0.48 | 63.01 ± 5.79 |
| Elastic | 87.50 ± 0.37 | 87.12 ± 0.57 | 86.69 ± 0.78 | 83.09 ± 0.79 | 65.87 ± 3.69 |
| FedVG (Ours) | 88.30 ± 0.66 | 87.83 ± 0.39 | 87.40 ± 0.74 | 83.10 ± 0.54 | **75.18 ± 1.36** |
| *DermaMNIST / ResNet-50* | | | | | |
| FedAvg | 83.49 ± 0.20 | 83.24 ± 0.42 | 83.38 ± 0.51 | 78.32 ± 0.50 | 70.58 ± 0.87 |
| FedAvgM | **85.06 ± 0.33** | **84.19 ± 0.38** | **83.99 ± 0.52** | 77.92 ± 0.56 | 53.18 ± 28.91 |
| FedProx | 83.61 ± 0.62 | 83.81 ± 0.39 | 83.32 ± 0.77 | **78.37 ± 0.66** | 70.70 ± 0.87 |
| Scaffold | 83.26 ± 0.21 | 83.59 ± 0.20 | 82.48 ± 0.59 | 74.02 ± 1.77 | 66.50 ± 2.76 |
| FedDyn | 75.44 ± 0.47 | 75.93 ± 0.44 | 76.54 ± 0.67 | 71.32 ± 0.81 | 66.42 ± 2.50 |
| Elastic | 82.66 ± 0.20 | 82.31 ± 0.62 | 81.79 ± 0.37 | 78.00 ± 0.38 | 70.22 ± 1.07 |
| FedVG (Ours) | 83.66 ± 0.31 | 83.45 ± 0.31 | 83.58 ± 0.55 | 77.82 ± 0.46 | **71.31 ± 0.71** |

show the class mappings from CIFAR-100 and STL-10 to CIFAR-10, respectively. After mapping, we applied balanced random sampling to select a fixed number of examples per class from the external datasets, ensuring a repre-

Table A.4. Test accuracy (%) ± standard deviation of federated learning methods on COVID19 with ViT-S/16 and DermaMNIST with ViT-B/16 models across different data heterogeneity levels (Dirichlet $\alpha$). **Bold** values indicate best accuracies and underlined values indicate second best accuracies.

| Methods | $\alpha = 100$ | $\alpha = 10$ | $\alpha = 1$ | $\alpha = 0.1$ | $\alpha = 0.05$ |
|---|---|---|---|---|---|
| *COVID19 / ViT-S/16* | | | | | |
| FedAvg | 88.38 ± 0.73 | 89.07 ± 0.49 | 89.63 ± 0.47 | 87.46 ± 0.72 | 82.30 ± 3.63 |
| FedAvgM | 88.01 ± 0.75 | 88.77 ± 0.50 | 89.14 ± 0.38 | 80.06 ± 8.58 | 51.66 ± 8.36 |
| FedProx | **88.40 ± 0.57** | **89.69 ± 0.12** | 89.32 ± 0.52 | **87.98 ± 1.06** | 81.27 ± 3.57 |
| Scaffold | 84.41 ± 4.02 | 84.80 ± 4.78 | 85.75 ± 4.33 | 68.33 ± 15.35 | 75.41 ± 9.81 |
| FedDyn | 77.61 ± 1.61 | 78.16 ± 1.64 | 76.72 ± 1.50 | 51.09 ± 8.63 | 42.42 ± 4.94 |
| Elastic | 88.37 ± 0.68 | 89.13 ± 0.27 | **89.66 ± 0.80** | 87.04 ± 1.22 | **85.06 ± 1.05** |
| FedVG (Ours) | 88.29 ± 0.33 | 89.06 ± 0.29 | 89.26 ± 0.45 | 87.47 ± 0.78 | 83.34 ± 1.53 |
| *DermaMNIST / ViT-B/16* | | | | | |
| FedAvg | 81.13 ± 0.27 | 81.31 ± 0.65 | 80.85 ± 0.50 | 77.79 ± 1.84 | 73.53 ± 2.40 |
| FedAvgM | 80.65 ± 0.73 | 81.08 ± 0.40 | 80.43 ± 1.39 | 71.74 ± 2.18 | 67.15 ± 0.32 |
| FedProx | 78.64 ± 6.28 | 80.85 ± 0.94 | 81.03 ± 0.31 | 78.50 ± 1.02 | 73.29 ± 3.79 |
| Scaffold | **81.43 ± 0.48** | 81.48 ± 0.42 | 81.41 ± 0.38 | 76.44 ± 1.47 | 75.36 ± 2.68 |
| FedDyn | 67.85 ± 1.45 | 66.87 ± 0.06 | 66.34 ± 0.76 | 66.67 ± 0.56 | 66.98 ± 0.04 |
| Elastic | 80.96 ± 0.35 | 79.44 ± 3.27 | 81.03 ± 0.60 | **79.48 ± 0.67** | 75.50 ± 1.41 |
| FedVG (Ours) | 80.46 ± 2.31 | **81.60 ± 0.26** | **81.46 ± 0.50** | 78.31 ± 0.99 | **76.20 ± 0.76** |

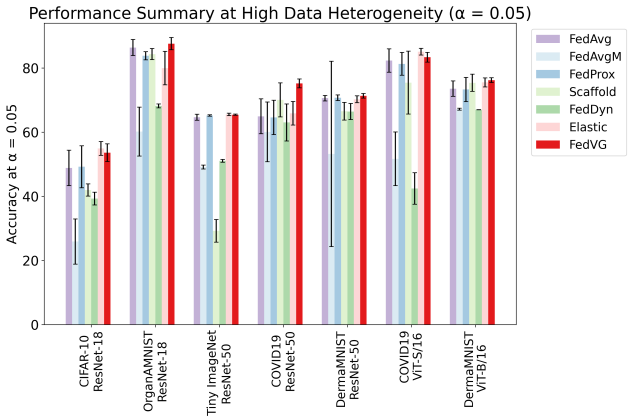



Figure A.5. Performance comparison of federated learning methods across datasets at high heterogeneity ($\alpha = 0.05$). The error bars denote standard deviation across five runs. At high heterogeneity, FedVG maintains both strong average performance and low variance compared to other methods.

sentative validation set with uniform class coverage. A total of 3,650 samples from CIFAR-100 and 4,950 samples from STL-10 were selected.

## A.6. FedVG Integration with FL Algorithms

Table A.8 presents numeric results for FedVG's integration with different federated learning methods. The experiments cover three datasets: CIFAR-10 and OrganAMNIST (us-

Table A.5. Mean accuracy comparison between FedVG and FedAWA across multiple datasets and model architectures under varying levels of data heterogeneity.

| $\alpha$ | FedAWA | FedVG |
|---|---|---|
| 0.05 | 72.61 | **73.23** |
| 0.1 | 78.41 | **78.63** |
| 1 | 83.32 | **83.63** |
| 10 | 83.94 | **84.30** |
| 100 | **84.33** | 84.30 |

Table A.6. Class mapping from STL-10 to CIFAR-10

| **CIFAR-10** | **STL-10** | **Count** |
|---|---|---|
| airplane | airplane | 550 |
| automobile | car | 550 |
| bird | bird | 550 |
| cat | cat | 550 |
| deer | deer | 550 |
| dog | dog | 550 |
| frog | — | — |
| horse | horse | 550 |
| ship | ship | 550 |
| truck | truck | 550 |

Table A.7. Class mapping from CIFAR-100 to CIFAR-10

| CIFAR-10 | CIFAR-100 | Count |
|---|---|---|
| airplane | — | — |
| automobile | bus, motorcycle | 531, 519 |
| bird | — | — |
| cat | leopard, lion, tiger | 520, 514, 516 |
| deer | — | — |
| dog | wolf, fox | 520, 530 |
| frog | — | — |
| horse | — | — |
| ship | — | — |
| truck | — | — |

ing ResNet-18), and COVID-19 (using ResNet-50), which are evaluated across a range of data heterogeneity settings. Each baseline FL algorithm, including FedAvg, FedAvgM, FedProx, Scaffold, FedDyn, and Elastic, is assessed both alone and combined with FedVG.

## A.7. Details of Ablation Analysis

### A.7.1. Analysis of Global Validation Set $D_{val}$

**Class Imbalance:** To simulate varying levels of class imbalance in the global validation set, we employ a controlled sampling strategy based on the imbalance ratio, $\rho \in (0, 1]$, where $\rho = 1$ indicates a balanced class distribution. To ensure a fair comparison, we fix the total size of the imbalanced validation set to half the size of the original validation set, i.e., $N' = \frac{1}{2}|D_{val}|$.

First we compute the unnormalized class proportion as:

$$p_i' = \rho^i, \quad i = 0, 1, \ldots, C-1. \tag{E.8}$$

These are then normalized to obtain valid class probabilities so that $\sum_{i=0}^{C-1} p_i = 1$.

$$p_i = \frac{\rho^i}{\sum_{j=0}^{C-1} \rho^j} \tag{E.9}$$

Finally, the number of validation samples allocated to each class is given by $s_i = \lfloor p_i \times N' \rfloor$, where, $\lfloor . \rfloor$ denotes the floor operation.

Figure A.6 shows examples of class distributions for $D_{val}$ of the CIFAR-10 dataset, as the imbalance ratio $\rho \to 0$.

**Size of Global Validation Set:** In addition to class imbalance, we also study the impact of the size of the global validation set on FedVG's performance on CIFAR-10 with the ResNet-18 architecture. To simulate different validation set sizes, we use stratified sampling to form subsets of the original validation dataset $D_{val}$ at varying proportions. As shown in Figure A.7, FedVG remains stable even when the global validation set is substantially reduced. In particular, with only 119 validation samples (0.025 fraction of $D_{val}$), FedVG still outperforms FedAvg by up to 3.23%.

### A.7.2. Evaluation of norm type

Eqn. 4 represents a general norm for converting layer-wise gradients into a mean gradient value, where $\|\cdot\|$ denotes a general vector or matrix norm. Common choices for $\|\cdot\|$ include $\ell_1$-norm, $\ell_2$-norm, and spectral norm (i.e., the largest singular value of the gradient matrix), etc.

Formally, for a vector $\mathbf{g} = (g_1, g_2, \ldots, g_d) \in \mathbb{R}^d$,

- The $\ell_1$-norm is defined as $\|\mathbf{g}\|_1 = \sum_{i=1}^{d} |g_i|$.
- The $\ell_2$-norm is defined as $\|\mathbf{g}\|_2 = \sqrt{\sum_{i=1}^{d} g_i^2}$.

For a matrix $G \in \mathbb{R}^{m \times n}$, representing the gradient of a layer,

- the spectral norm $\|G\|_\sigma$ is defined as the largest singular value of $G$, which corresponds to $\|G\|_\sigma = \sigma_{\max}(G)$
- where, $\sigma_{\max}(G)$ is the maximum singular value obtained from the singular value decomposition (SVD) of$G$.

When calculating the $\ell_1$-norm and $\ell_2$-norm, matrices are first flattened into vectors. For the spectral norm, only layer gradients with two or more dimensions are considered and reshaped into matrices for singular value decomposition.

In addition, we define the *delta norm*, which is computed as the norm of parameter updates (i.e., the difference between client models and the global model) rather than the layer gradients. For *delta norm*, Eqn. 4 becomes

$$\bar{G}_k = \frac{1}{L} \sum_{\ell=1}^{L} \left\| \theta_g^l - \theta_k^l \right\|. \tag{E.10}$$

Figure A.8 shows the client setup for these experiments. Here, 9 clients contain the same number of elements in their local dataset, with a concentration coefficient $\alpha = 0.05$. The tenth client (client 1) is made to be perfectly homogeneous. Figure 8 then shows the FedVG weights assigned to each client at each round of federated training.

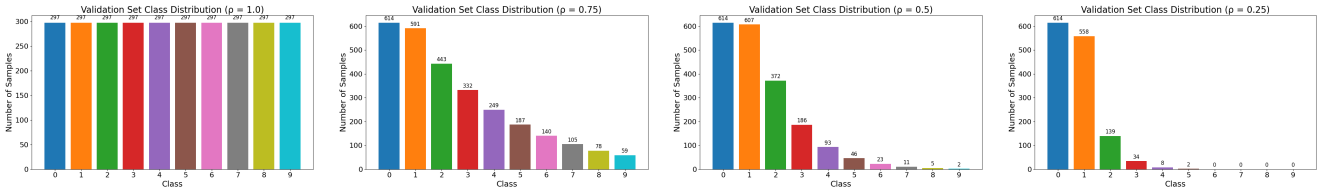
### A.7.3. Aggregation granularity

To analyze the impact of aggregation granularity, we evaluate three strategies: modelwise, layerwise, and blockwise aggregation.

In the modelwise setting (the default FedVG configuration), the client's mean gradient magnitude is computed by averaging gradient norms across all $L$ layers

$$\bar{G}_k = \frac{1}{L} \sum_{\ell=1}^{L} \left\| \nabla_{\theta_k^{(\ell)}} \mathcal{L}_{\text{val}} \right\|. \tag{E.11}$$

and the corresponding raw score is




(a) $\rho = 1$ (b) $\rho = 0.75$ (c) $\rho = 0.5$ (d) $\rho = 0.25$

Figure A.6. Visualization of class imbalance of $D_{val}$ at various values of $\rho$.

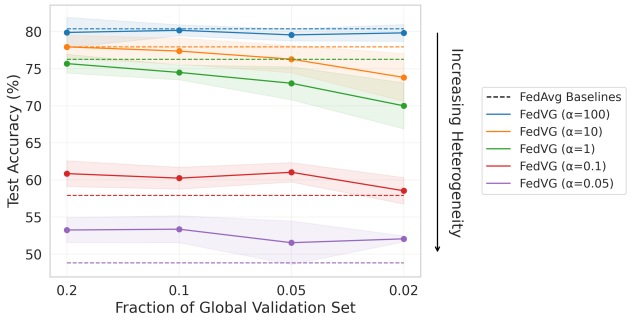



Figure A.7. FedVG performance across different fractions of the global validation set.

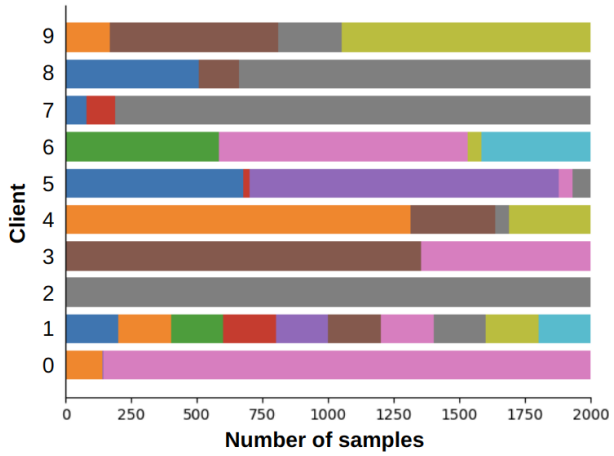



Figure A.8. Visualization of sample distributions used in the norm-type ablation study. Here, client one is configured to be perfectly balanced.

$$\hat{s_k} = \frac{1}{\bar{G}_k + \epsilon} \tag{E.12}$$

In the layerwise variant, we compute the raw score ($\hat{s_k}^{(l)}$) and normalized score ($s_k^{(l)}$) for each layer $l$ separately. The raw and normalized scores are given by:

$$G_k^{(l)} = \left\| \nabla_{\theta_k^{(\ell)}} \mathcal{L}_{\text{val}} \right\| \tag{E.13}$$

$$\hat{s_k}^{(l)} = \frac{1}{G_k^{(l)} + \epsilon} \tag{E.14}$$

$$s_k^{(l)} = \frac{\hat{s_k}^{(l)}}{\sum_{j=1}^{K} \hat{s_j}^{(l)}} \tag{E.15}$$

The aggregation step is then performed for each layer:

$$\theta_g^{(l)} \leftarrow \theta_g^{(l)} - \sum_{k=1}^{K} s_k^{(l)} (\theta_g^{(l)} - \theta_k^{(l)}), \quad \forall\, l = 1, 2, \ldots, L \tag{E.16}$$

In the blockwise variant, we follow a similar approach to modelwise aggregation, except the averaging is performed over layers within each block rather than the entire model. For ResNet models, each residual block is treated as a block; for ViT-based models, each transformer encoder block is treated as a block; and in both cases, the classifier head is considered a separate block. The model $\theta_k$ can be decomposed into $B$ blocks,

$$\theta_k = \{\theta_k^{(1)}, \theta_k^{(2)}, \ldots, \theta_k^{(B)}\} \tag{E.17}$$

Let $|b|$ denote the number of layers in block $b$. The block-level score computation is:

$$\bar{G}_k^{(b)} = \frac{1}{|b|} \sum_{\ell \in b} \left\| \nabla_{\theta_k^{(\ell)}} \mathcal{L}_{\text{val}} \right\| \tag{E.18}$$

$$\hat{s_k}^{(b)} = \frac{1}{G_k^{(b)} + \epsilon} \tag{E.19}$$

$$s_k^{(b)} = \frac{\hat{s_k}^{(b)}}{\sum_{j=1}^{K} \hat{s_j}^{(b)}} \tag{E.20}$$

The aggregation is then applied at the block level:

$$\theta_g^{(b)} \leftarrow \theta_g^{(b)} - \sum_{k=1}^{K} s_k^{(b)} (\theta_g^{(b)} - \theta_k^{(b)}), \quad \forall\, b = 1, 2, \ldots, B. \tag{E.21}$$

Table A.9 reports the detailed performance of the three FedVG variants (Modelwise, Layerwise, and Blockwise), across different dataset–model combinations and heterogeneity levels ($\alpha \in 100, 10, 1, 0.1, 0.05$).

Table A.8. Performance comparison of federated learning methods and their integration with FedVG on CIFAR-10, OrganAMNIST, and COVID-19 datasets using ResNet-18 and ResNet-50 models across data heterogeneity levels ($\alpha$). **Bold** values indicate best accuracies and underlined values indicate second best accuracies.

| Method | $\alpha = 100$ | $\alpha = 10$ | $\alpha = 1$ | $\alpha = 0.1$ | $\alpha = 0.05$ |
|---|---|---|---|---|---|
| *CIFAR-10 / ResNet-18* | | | | | |
| FedAvg | 80.37 ± 0.89 | 77.95 ± 0.43 | **76.26 ± 1.65** | 57.93 ± 1.82 | 48.83 ± 5.53 |
| FedVG | 80.74 ± 0.79 | **79.65 ± 0.52** | 75.69 ± 2.27 | **61.06 ± 0.34** | **53.58 ± 2.78** |
| FedAvg + FedVG | **80.92 ± 0.79** | 78.79 ± 0.56 | 74.96 ± 2.55 | 60.60 ± 2.26 | 52.49 ± 3.83 |
| FedAvgM | 71.09 ± 5.16 | 56.64 ± 10.75 | **51.87 ± 11.40** | 34.18 ± 4.38 | 25.87 ± 7.04 |
| FedAvgM + FedVG | **77.87 ± 1.98** | **63.05 ± 12.55** | 39.51 ± 17.02 | **42.63 ± 7.25** | **33.25 ± 3.62** |
| FedProx | 79.92 ± 1.50 | 77.39 ± 1.32 | **76.73 ± 0.63** | 55.22 ± 5.56 | 49.19 ± 6.55 |
| FedProx + FedVG | **81.16 ± 0.21** | **79.10 ± 1.85** | 74.60 ± 1.75 | **60.18 ± 2.18** | **52.67 ± 3.12** |
| Scaffold | **81.05 ± 0.16** | 80.39 ± 0.15 | **79.25 ± 0.23** | 56.15 ± 2.65 | **41.94 ± 1.91** |
| Scaffold + FedVG | 80.87 ± 0.15 | **80.42 ± 0.15** | 79.22 ± 0.18 | **56.59 ± 1.28** | 39.94 ± 3.13 |
| FedDyn | 66.37 ± 0.33 | 65.90 ± 0.36 | 62.37 ± 0.29 | **41.66 ± 1.22** | **39.26 ± 1.99** |
| FedDyn + FedVG | **68.62 ± 0.32** | **68.27 ± 0.39** | **63.34 ± 0.52** | 37.25 ± 0.32 | 32.59 ± 0.74 |
| Elastic | 79.45 ± 0.27 | 76.55 ± 1.88 | 75.69 ± 0.89 | 55.86 ± 4.26 | **54.92 ± 2.16** |
| Elastic + FedVG | **79.95 ± 0.23** | **79.34 ± 0.62** | **76.46 ± 0.97** | **59.09 ± 4.07** | 51.75 ± 4.49 |
| *OrganAMNIST / ResNet-18* | | | | | |
| FedAvg | 98.46 ± 0.16 | 98.87 ± 0.21 | 98.47 ± 0.24 | 93.50 ± 1.22 | 86.37 ± 2.49 |
| FedVG | **99.41 ± 0.08** | **99.42 ± 0.02** | **99.12 ± 0.10** | **94.72 ± 0.68** | **87.57 ± 1.91** |
| FedAvg + FedVG | 98.83 ± 0.29 | 99.22 ± 0.11 | 98.76 ± 0.30 | 94.35 ± 1.24 | 87.09 ± 1.25 |
| FedAvgM | 87.14 ± 4.56 | 93.43 ± 4.13 | 94.00 ± 4.76 | 78.95 ± 8.14 | 60.15 ± 7.61 |
| FedAvgM + FedVG | **99.20 ± 0.07** | **99.52 ± 0.06** | **98.77 ± 0.70** | **88.87 ± 5.06** | **77.01 ± 7.10** |
| FedProx | 98.44 ± 0.21 | 98.68 ± 0.32 | 98.60 ± 0.20 | 93.19 ± 1.63 | 83.80 ± 1.25 |
| FedProx + FedVG | **99.44 ± 0.05** | **99.43 ± 0.05** | **99.02 ± 0.17** | **94.97 ± 0.37** | **88.00 ± 1.24** |
| Scaffold | **99.22 ± 0.06** | **99.17 ± 0.06** | **98.85 ± 0.06** | 93.11 ± 0.62 | **84.32 ± 1.72** |
| Scaffold + FedVG | 99.14 ± 0.10 | **99.17 ± 0.07** | 98.80 ± 0.06 | **93.19 ± 0.27** | 83.19 ± 4.53 |
| FedDyn | 97.26 ± 0.17 | 97.13 ± 0.07 | 95.60 ± 0.15 | **78.61 ± 1.01** | **68.17 ± 0.62** |
| FedDyn + FedVG | **97.79 ± 0.08** | **97.78 ± 0.07** | **96.24 ± 0.23** | 77.25 ± 1.58 | 63.31 ± 1.88 |
| Elastic | 99.12 ± 0.11 | 98.78 ± 0.18 | 97.94 ± 0.27 | 88.46 ± 3.70 | 79.96 ± 5.20 |
| Elastic + FedVG | **99.31 ± 0.05** | **99.33 ± 0.08** | **99.00 ± 0.08** | **92.98 ± 1.39** | **87.95 ± 1.00** |
| *COVID-19 / ResNet-50* | | | | | |
| FedAvg | 88.23 ± 0.23 | 87.88 ± 0.38 | **87.91 ± 0.37** | **84.34 ± 0.67** | 64.92 ± 5.45 |
| FedVG | **88.30 ± 0.66** | 87.83 ± 0.39 | 87.40 ± 0.74 | 83.10 ± 0.54 | **75.18 ± 1.36** |
| FedAvg + FedVG | 88.26 ± 0.53 | **88.38 ± 0.54** | 87.88 ± 0.59 | 83.62 ± 0.83 | 68.39 ± 4.12 |
| FedAvgM | **90.16 ± 0.23** | **90.23 ± 0.54** | 90.00 ± 0.33 | 85.54 ± 0.87 | 60.07 ± 9.31 |
| FedAvgM + FedVG | 89.96 ± 0.46 | 89.88 ± 0.55 | **90.02 ± 0.25** | **88.03 ± 0.52** | **72.66 ± 4.18** |
| FedProx | 87.93 ± 0.55 | 87.91 ± 0.37 | **87.53 ± 0.47** | **84.06 ± 1.06** | 64.58 ± 5.34 |
| FedProx + FedVG | **88.22 ± 0.22** | **88.39 ± 0.46** | 87.32 ± 0.53 | 83.43 ± 0.80 | **72.55 ± 2.54** |
| Scaffold | **88.70 ± 0.30** | 88.67 ± 0.55 | 88.36 ± 0.26 | **84.47 ± 1.06** | 70.05 ± 5.29 |
| Scaffold + FedVG | 88.63 ± 0.28 | **88.88 ± 0.40** | **88.64 ± 0.28** | 83.77 ± 0.93 | **71.62 ± 3.74** |
| FedDyn | 84.48 ± 0.48 | **84.51 ± 0.33** | 84.28 ± 0.75 | **78.49 ± 0.48** | 63.01 ± 5.79 |
| FedDyn + FedVG | **88.30 ± 0.66** | 84.03 ± 0.44 | **84.31 ± 0.74** | 76.32 ± 3.57 | **66.99 ± 3.98** |
| Elastic | **87.50 ± 0.37** | **87.12 ± 0.57** | 86.69 ± 0.78 | **83.09 ± 0.79** | 65.87 ± 3.69 |
| Elastic + FedVG | 86.99 ± 0.44 | 86.71 ± 0.44 | **86.95 ± 0.75** | 82.70 ± 0.68 | **75.24 ± 2.12** |

Table A.9. Comparison of FedVG variants (Modelwise, Layerwise, and Blockwise) across datasets, models, and heterogeneity levels ($\alpha$ values). **Bold** values indicate best accuracies.

| Method | $\alpha = 100$ | $\alpha = 10$ | $\alpha = 1$ | $\alpha = 0.1$ | $\alpha = 0.05$ |
|---|---|---|---|---|---|
| *CIFAR-10 / ResNet-18* | | | | | |
| FedVG (Modelwise) | **80.74 ± 0.79** | **79.65 ± 0.52** | **75.69 ± 2.27** | **61.06 ± 0.34** | **53.58 ± 2.78** |
| FedVG (Layerwise) | 77.87 ± 2.79 | 73.33 ± 5.30 | 74.03 ± 2.41 | 59.43 ± 2.24 | 50.25 ± 1.80 |
| FedVG (Blockwise) | 79.33 ± 1.85 | 71.15 ± 8.79 | 71.85 ± 2.96 | 58.06 ± 3.87 | 50.30 ± 2.95 |
| *OrganAMNIST / ResNet-18* | | | | | |
| FedVG (Modelwise) | **99.41 ± 0.08** | **99.42 ± 0.02** | **99.12 ± 0.10** | **94.72 ± 0.6** | **87.57 ± 1.91** |
| FedVG (Layerwise) | 99.33 ± 0.09 | 99.30 ± 0.29 | 98.94 ± 0.17 | 94.30 ± 1.86 | 87.44 ± 0.99 |
| FedVG (Blockwise) | 99.29 ± 0.06 | 99.39 ± 0.06 | 99.02 ± 0.16 | 92.83 ± 2.16 | 87.02 ± 1.42 |
| *Tiny ImageNet / ResNet-50* | | | | | |
| FedVG (Modelwise) | 69.21 ± 0.10 | 69.09 ± 0.05 | 68.93 ± 0.18 | 67.21 ± 0.27 | **65.42 ± 0.18** |
| FedVG (Layerwise) | 69.23 ± 0.15 | **69.12 ± 0.09** | 68.79 ± 0.45 | 66.85 ± 0.56 | 64.88 ± 0.99 |
| FedVG (Blockwise) | **69.31 ± 0.15** | **69.12 ± 0.11** | **69.02 ± 0.13** | **67.27 ± 0.19** | 65.09 ± 0.79 |
| *COVID-19 / ResNet-50* | | | | | |
| FedVG (Modelwise) | 88.30 ± 0.66 | 87.83 ± 0.39 | 87.40 ± 0.74 | 83.10 ± 0.54 | **75.18 ± 1.36** |
| FedVG (Layerwise) | **87.88 ± 0.51** | **87.88 ± 0.36** | 87.42 ± 0.54 | 83.71 ± 0.78 | 73.44 ± 1.74 |
| FedVG (Blockwise) | 88.38 ± 0.59 | 88.23 ± 0.42 | **87.75 ± 0.54** | **83.54 ± 0.90** | 74.23 ± 1.73 |
| *DermaMNIST / ResNet-50* | | | | | |
| FedVG (Modelwise) | **83.66 ± 0.31** | **83.45 ± 0.31** | 83.58 ± 0.55 | 77.82 ± 0.46 | 71.31 ± 0.71 |
| FedVG (Layerwise) | 83.46 ± 0.27 | 83.08 ± 0.40 | 83.19 ± 0.51 | **77.83 ± 0.48** | 71.22 ± 0.72 |
| FedVG (Blockwise) | 83.53 ± 0.65 | 83.21 ± 0.17 | **83.79 ± 0.18** | 77.78 ± 0.54 | **71.48 ± 1.00** |
| *COVID-19 / ViT-S/16* | | | | | |
| FedVG (Modelwise) | 88.29 ± 0.33 | **89.06 ± 0.29** | 89.26 ± 0.45 | **87.47 ± 0.78** | 83.34 ± 1.53 |
| FedVG (Layerwise) | **88.41 ± 0.68** | 88.66 ± 0.73 | 89.57 ± 0.32 | 87.17 ± 0.58 | 83.45 ± 1.92 |
| FedVG (Blockwise) | 87.98 ± 0.69 | 88.98 ± 0.20 | **89.66 ± 0.63** | 87.01 ± 0.61 | **83.90 ± 1.74** |
| *DermaMNIST / ViT-B/16* | | | | | |
| FedVG (Modelwise) | 80.46 ± 2.31 | **81.60 ± 0.26** | 81.46 ± 0.50 | 78.31 ± 0.99 | 76.20 ± 0.76 |
| FedVG (Layerwise) | **80.72 ± 1.60** | 81.35 ± 0.43 | **81.47 ± 0.26** | **79.25 ± 0.80** | 73.92 ± 2.51 |
| FedVG (Blockwise) | 79.70 ± 2.87 | 81.20 ± 0.56 | 79.63 ± 2.89 | 79.05 ± 0.85 | **76.29 ± 0.67** |