Navigating Heterogeneity and Privacy in One-Shot Federated Learning with Diffusion Models – Supplementary Material

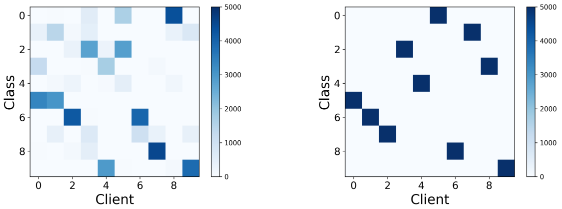
Matías Mendieta, Guangyu Sun, Chen Chen
Center for Research in Computer Vision, University of Central Florida, USA
{matias.mendieta, guangyu}@ucf.edu; chen.chen@crcv.ucf.edu

1. Additional Training Details

The FashionMNIST dataset is an alternative to the original MNIST dataset, providing a more challenging task by replacing the handwritten digits with grayscale images of various fashion items. The dataset consists of 60,000 training images and 10,000 test images. The PathMNIST dataset is a medical dataset of colon pathology images in RGB, with a training set of 89,996 images and a test set containing 7,180 images with 9 classes. The CIFAR-10 dataset consists of 60,000 color images equally distributed into ten different classes. The dataset is composed of a training set containing 50,000 images and a test set comprising of 10,000 images. CIFAR-10 is natively sized at 32×32 pixels. We upsample FashionMNIST and PathMNIST from 28×28 to 32×32. A visualization of the dataset partitioning across clients is shown in Figure 1.

We train with a batch size of 128 for all methods and use the AdamW optimizer. For local (and global training were applicable), we searched learning rates from [$3e^{-3}$, $1e^{-3}$, $3e^{-4}$, $1e^{-4}$] for each method using the CIFAR-10 dataset to find the optimal settings. We employed a ResNet16 architecture for the global model of all methods to ensure a fair comparison. For DP experiments, we set the max gradient norm clipping threshold to 1.0 for all experiments and methods. In accordance with the recommendations of the Opacus [5] library, we employ their Poisson batch sampling to ensure privacy guarantees.

As mentioned in Section 3.1 of the main paper, our DM is a basic U-Net structure with residual blocks [3, 4] and class-conditioning. Specifically, our U-Net has three downsampling stages (1/8 total downsampling) and three upsampling stages, each with residual convolutional blocks. A learnable embedding for time step and class conditioning is concatenated as additional input channels. For FedDiff$_S$, we halve the number of channels per block to reduce model size. For sampling at the server, we perform 1000 iterations as in [3] to generate each batch. The total number of generated samples is set equal to the size of the original dataset. Code available at https://github.com/mmendiet/FedDiff.



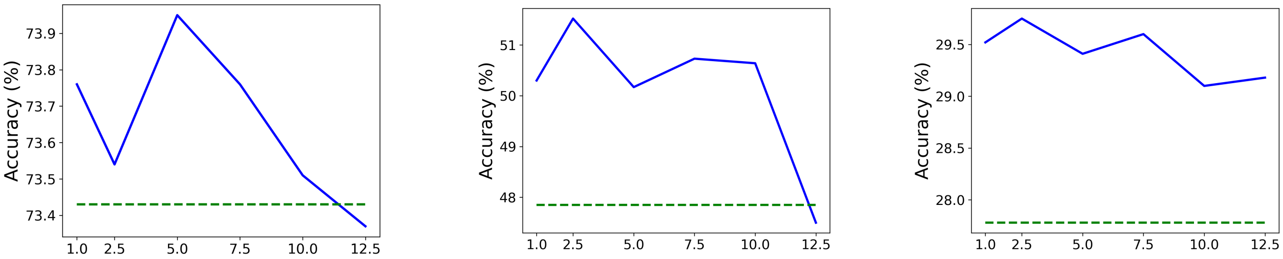
(a) $\alpha = 0.1$ (b) $\alpha = 0.001$

Figure 1. $Dir(\alpha)$ data partitioning for 10 clients on CIFAR-10. We show moderate ($\alpha = 0.1$) to severe ($\alpha = 0.01$) data heterogeneity levels. Data heterogeneity poses a significant challenge for many one-shot FL methods, as reconciling various models trained on widely different distributions is non-trivial. Our FedDiff approach rather trains diffusion models on the simple client distributions, which can then generate useful synthetic data for training global models.

2. Communication and Server-side Operations

All methods primarily involve transmitting the model weights to the server, and this is done a single time. Additionally, specific information is sent. FedAvg and DENSE transmit the number of samples and label space, while FedCVAE and FedDiff send the number of samples per label. Thus, communication cost is mainly determined by the model size and number of clients. To ensure fairness between generative and discriminative methods, we select models with similar parameters and FLOPs, as shown in Table 3, maintaining comparable communication and computation costs. We describe details of our DM architectures in Section 1. As we also show in our experiments with FedDiff$_S$ in Table 3 of the main paper, we can adjust the size of the model to meet communication or compute needs and still provide exceptional performance.

Once the models are on the server, different operations are required for each method. DENSE, FedCVAE, and FedDiff require server-side generation and training, producing



(a) FashionMNIST

(b) PathMNSIT

(c) CIFAR-10

Figure 2. Ablation study of γ in FMF under the $\epsilon = 10$ setting. The accuracy of FedDiff is in **green** and FedDiff+FMF for various γ in **blue**. Generally, data filtering within the range of 1% to 10% produces positive outcomes, resulting in improved performance, with approximately 5% serving as an effective default choice. We plot the mean across three runs with different seeds for each setting.

a dataset the same size as the original (e.g., 50k for CIFAR-10) and training a global model for 200 epochs. Server-side operation times on a single A5000 GPU for CIFAR-10 with 10 clients are 3.36, 1.76, and 2.17 hours for DENSE, FedCVAE, and FedDiff, respectively. DENSE takes the longest due to nested training of a GAN and the global model. Nonetheless, server computation is not a primary concern in FL, as it is not constrained by the same resource limitations as client devices.

3. Sampling Steps for Generation Ablation

Diffusion models generate images through consecutive denoising steps, making the number of iterations per image a design choice when generating synthetic datasets at the server. Table 1 provides insights into the impact of adjusting this parameter. Generally, increasing the number of steps enhances data quality and improves the final global model performance, particularly for the more difficult datasets. However, if generation time on the server is a concern, this parameter can be reduced, or increased when prioritizing data quality and model accuracy.

Table 1. Ablation on number of diffusion steps (S) used per image when generating the global synthetic dataset. Final global model accuracy is reported using our FedDiff approach under the default setting of $\alpha = 0.01$ and $C = 10$. The 1000-step setting is employed across all other experiments in the paper, as it is the standard practice in DDPM [3].

Dataset	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$
FashionMNIST	86.49±0.23	86.63±0.38	86.81±0.54	86.75±0.17
PathMNIST	69.85±1.15	70.13±1.61	70.61±1.37	71.44±1.83
CIFAR-10	55.89±1.93	56.13±1.74	56.57±2.42	57.74±2.86

4. FMF γ Ablation

In Figure 2, we present the outcomes obtained using FedDiff+FMF under $\epsilon = 10$ across a range of γ values, encompassing data filtering percentages spanning from 1% to 12%. Our findings indicate that, in general, data filtering within the 1% to 10% range yields favorable results and leads to performance enhancements, with around 5% being a great default. Interestingly, the degree of improvement provided by FMF becomes more pronounced and consistent as the dataset becomes more challenging. This phenomenon aligns with the anticipated trends, as more intricate datasets inherently pose a greater challenge, making it less likely for the generators to consistently produce high-quality samples. Consequently, the need for data filtering becomes more pronounced in such scenarios to enhance sample quality. This trend is also favorable since it addresses the specific need for improvement, especially in cases where performance is suboptimal and the challenges are more pronounced.

5. Discussions, Limitations and Broader Impact

Model Heterogeneity. In real FL systems, model heterogeneity may often occur [2,6]. For instance, some clients may have architecture variations in their models or have smaller or larger models depending on their computing capabilities. Therefore, clients may have different architectures of similar generation capability, or even differing capabilities depending on the requirements of each client. Our approach allows for flexibility to accommodate such system diversity across clients. In FedDiff, we generate data from the client models and employ that synthetic data for global training, and therefore can leverage varying models without the worry of reconciling the weights themselves.

Limitations and Broader Impact. One downside of our method is that the generated data, particularly under DP constraints, still lacks in quality and effectiveness for global model training versus using true data. For instance, with DP on CIFAR-10 as shown in Figure 4 in the main paper, the data loses a substantial amount of structure. An interesting direction for future work would be to study how to further improve the quality of the generated data and its usefulness for global model training while maintaining privacy. For instance, as differential privacy algorithms improve, FedDiff and the generated data quality will likewise benefit. Additionally, potentially leveraging prior information could allow the models to focus on task-relevant features, excluding irrelevant ones. By training a separate model to identify important features and then learning to generate images based only on these features with a DP secure model, we could potentially simplify the information needed for the generation process. This would likely lead to quicker convergence, allowing for better learning in DP settings with the same privacy budget and ultimately improving the privacy-utility trade-off of the generated samples.

Looking at the broader impact of our work, FL depends on the diversity of data contributed by different participants. If biases exist in the local datasets, they can be propagated and amplified during the model training process. This could lead to unintended algorithmic biases and discrimination in the resulting models. Ensuring diversity and fairness in the data used for FL is an important research direction to mitigate this risk and promote equitable outcomes [1], particularly in the highly data heterogeneous environments explored in this work. Furthermore, as we have discussed throughout our paper, the privacy of client data is important in FL. To mitigate risks in this regard, we take many precautions to preserve privacy of the clients participated in the FL process though the use of DP, and operating within the one-shot setting to reduce the chance of eavesdropping.

References

[1] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *CoRR*, abs/2012.02447, 2020. 3

[2] Clare Elizabeth Heinbaugh, Emilio Luz-Ricca, and Huajie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. 1, 2

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[5] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021. 1

[6] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. *Advances in Neural Information Processing Systems*, 35:21414–21428, 2022. 2