# Efficient Model Personalization in Federated Learning via Client-Specific Prompt Generation Supplementary Material

Fu-En Yang[1,2] Chien-Yi Wang[2] Yu-Chiang Frank Wang[1,2]
[1]National Taiwan University [2]NVIDIA
{f07942077, ycwang}@ntu.edu.tw, chienyiw@nvidia.com

## A. Dataset Statistics

### A.1. Office-Caltech10

Office-Caltech10 dataset [17, 6] is comprised of four domains, including *Amazon*, *Caltech*, *DSLR*, and *Webcam*, and ten semantic categories, *backpack*, *bike*, *calculator*, *headphones*, *keyboard*, *laptop computer*, *monitor*, *mouse*, *mug*, and *projector*. The domain shifts mainly come from different camera devices or different background environments. The sampled images are provided in Fig. 1.

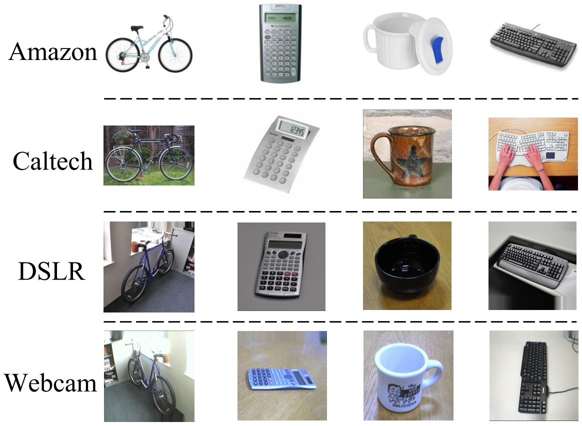


Figure 1. Sampled images from Office-Caltech10 [17, 6].

### A.2. DomainNet

Following the previous FL work FedBN [10], we construct a subset by selecting the top ten most frequent classes from the original DomainNet [13] for our experiments. The top ten frequent semantic categories contain *bird*, *feather*, *headphones*, *ice cream*, *teapot*, *tiger*, *whale*, *windmill*, *wine glass*, and *zebra*. In this dataset, images with the same type of style (*e.g., clipart* or *sketch*) form a domain. We provide the sampled images in Fig. 2.

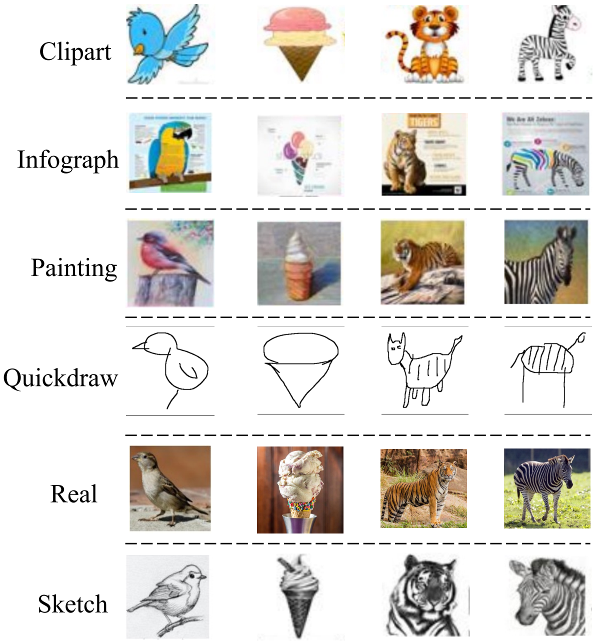


Figure 2. Sampled images from DomainNet [10, 13].

### A.3. Dermoscopic-FL

Dermoscopic-FL dataset [3] contains three types of skin lesions, *Nevus*, *Benign Keratosis*, and *Melanoma*, distributed among four data sites (*i.e., A, B, C*, and *D*) as summarized in Table 1. Data sites *A*, *B*, and *C* are collected from HAM10K [18] while data site *D* is collected from MSK [4]. In Fig. 3, we show sampled images from different data sites. The domain shifts among all data sites are caused by the use of different imaging devices.

## B. Further Analysis of Our pFedPG

### B.1. Impact of data size at clients

In Fig. 4, we analyze the impact of data size on clients compared with the baselines of FedAvg [15] and FedVPT

Table 1. Dataset statistics of Dermoscopic-FL dataset [3], which includes three types of skin lesions distributed among four data sites (*i.e.,* A, B, C, and D)

| Category | A | B | C | D |
|---|---|---|---|---|
| Nevus | 1,832 | 3,720 | 803 | 1,372 |
| Benign Keratosis | 475 | 124 | 490 | 254 |
| Melanoma | 680 | 24 | 342 | 374 |
| Total Images | 2,987 | 3,868 | 1,635 | 2,000 |

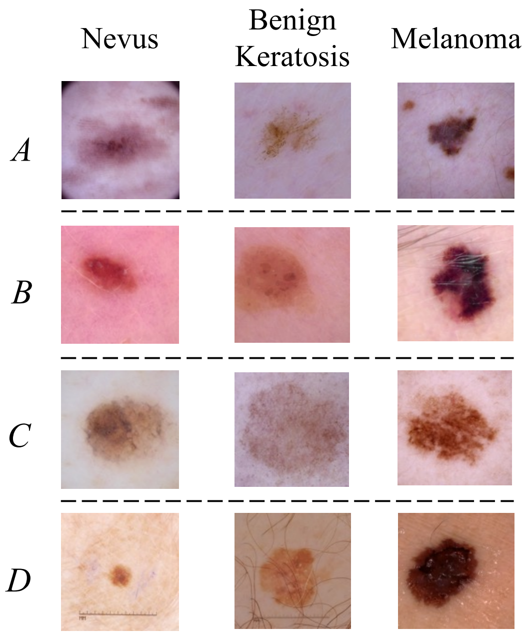
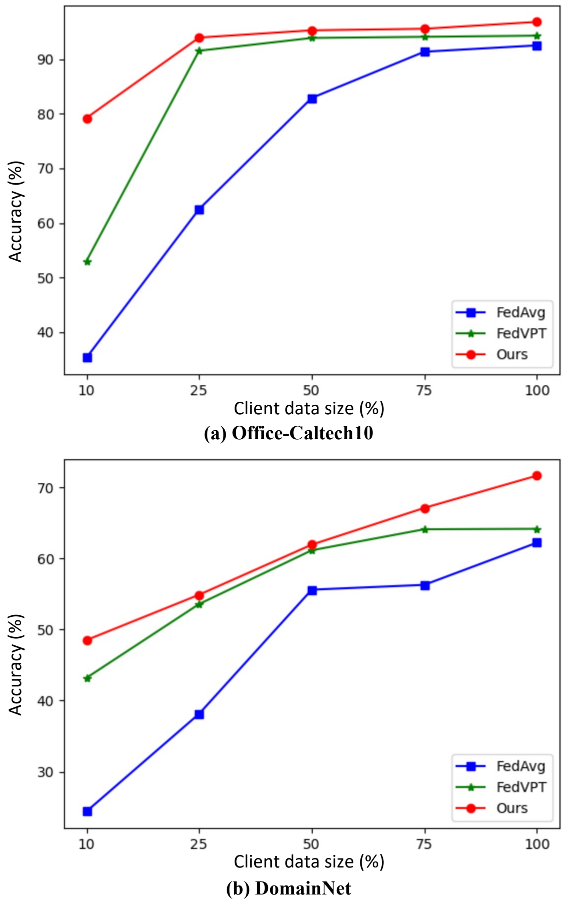


Figure 3. Sampled images from Dermoscopic-FL dataset [3].



**(a) Office-Caltech10**

**(b) DomainNet**

Figure 4. Impact of client data size on (a) Office-Caltech10 and (b) DomainNet datasets.

on the Office-Caltech10 and DomainNet datasets. As we can observe from Fig. 4, FedAvg [15] performed significantly inferior with prompt-based methods (*i.e.,* FedVPT and ours), especially when the client only contains limited data (*e.g.,* 10 % client data). This is due to the fact that updating entire network parameters of a large-scale model is prone to overfitting. In addition, our proposed pFedPG outperforms FedVPT at every data size, confirming the ability of our method to tackle heterogeneous local data distributions even in the few-data regime. From the above results, the robustness of our pFedPG is successfully verified.

## B.2. Impact of different pre-training models

In this section, we explore the ability of our proposed pFedPG framework to incorporate different types of pre-trained models. In addition to applying a ViT-B [5] pre-trained from ImageNet-21K in a fully supervised manner, we further evaluate the performance when the backbone is replaced by MAE [8] and MoCo v3 [2], both of which are trained in a self-supervised fashion. To be more specific, MAE [8] is trained using the masked image model objective, while MoCo v3 [2] is learned by contrastive loss. As shown in Table 2, we can observe that our proposed pFedPG outperforms the baselines of FedAvg and FedVPT regardless of the choice of pre-trained foundation models. From the above results, we verify that our pFedPG is generally applicable to various types of pre-trained foundation models. Furthermore, these findings confirm that when data heterogeneity exists across all clients, simply averaging whole model parameters or prompts from clients significantly diverges the aggregated model or prompts from each local distribution. Instead, our proposed approach leverage cross-client optimization directions to generate personalized prompts, which enable efficient model personalization. With the above results, the robustness and effectiveness of our pFedPG are successfully confirmed.

## B.3. Impact of different backbone architectures

In addition to adopting Vision Transformer (ViT) as our backbone architecture, we additionally consider hierarchical Transformers (*i.e.,* Swin Transformer [11]), which em-

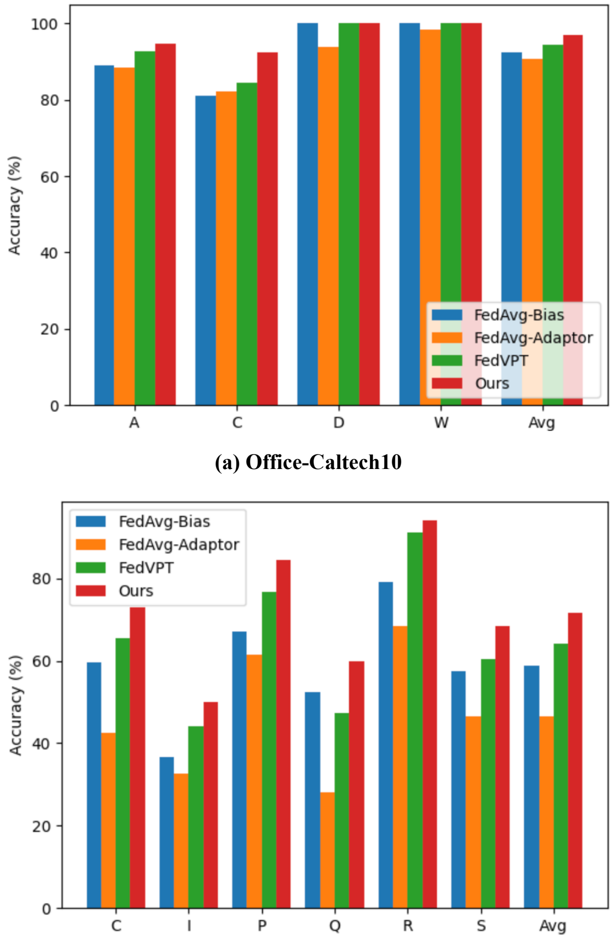
Table 2. Analysis of different pre-training methods and model backbones on benchmark datasets.

| Method | Pre-training | Backbone | Office-Caltech10 | DomainNet | CIFAR-10 | CIFAR-100 |
|---|---|---|---|---|---|---|
| FedAvg [15] | Supervised (21K) [5] | ViT-B | 92.51 | 62.21 | 79.79 | 51.37 |
|  | Supervised (21K) [11] | Swin-B | 53.31 | 50.64 | 67.54 | 34.63 |
|  | MAE (1K) [8] | ViT-B | 45.36 | 19.52 | 72.04 | 36.33 |
|  | MoCo v3 (1K) [2] | ViT-B | 91.23 | 63.99 | 75.32 | 44.33 |
| FedVPT | Supervised (21K) [5] | ViT-B | 94.29 | 64.16 | 85.11 | 45.26 |
|  | Supervised (21K) [11] | Swin-B | 57.49 | 57.23 | 95.66 | 80.91 |
|  | MAE (1K) [8] | ViT-B | 13.88 | 14.06 | 76.75 | 26.30 |
|  | MoCo v3 (1K) [2] | ViT-B | 93.87 | 63.09 | 85.56 | 45.24 |
| pFedPG (Ours) | Supervised (21K) [5] | ViT-B | 96.81 | 71.64 | 87.57 | 55.91 |
|  | Supervised (21K) [11] | Swin-B | 67.30 | 69.97 | 96.02 | 82.02 |
|  | MAE (1K) [8] | ViT-B | 53.90 | 33.39 | 82.26 | 39.29 |
|  | MoCo v3 (1K) [2] | ViT-B | 94.11 | 64.02 | 86.67 | 47.74 |

ploys multi-scale attention into locally shifted windows and aggregate image patch embeddings at deeper layers. Following VPT [9], prompts at each client are inserted within the local windows and are ignored during patch embedding aggregation. In Table 2, we conduct quantitative comparisons with baselines of FedAvg and FedVPT using Swin-B [11] pre-trained on ImageNet-21K as the backbone architecture. As we can observe in Table 2, our pFedPG is able to achieve generally preferable performances regardless of the backbone choice. The results summarized in Table 2 successfully confirm the robustness of our proposed pFedPG with various pre-training models and backbone architectures for tackling challenging heterogeneous federated learning settings.

## B.4. Additional comparisons with parameter-efficient tuning methods

In Fig. 5, we further compare our pFedPG with other types of parameter-efficient fine-tuning methods, including Bias [1] and Adaptor [14], on Office-Caltech10 and DomainNet datasets. Specifically, bias-tuning [1] only tunes the bias term while keeping the remaining model parameters frozen. On the other hand, adaptor-tuning [14] additionally inserts a few parameters (denoted as *adaptors*) inside the frozen backbone. We conduct baselines of FedAvg-Bias and FedAvg-Adaptor by integrating bias-tuning and adaptor-tuning into FedAvg [12] frameworks using ViT-B as backbones. From the results in Fig. 5, we observe that methods based on prompt-tuning (*i.e.,* FedVPT and ours) consistently perform superiorly against FedAvg-Bias and FedAvg-Adaptor on both benchmark datasets. The above results demonstrate the effectiveness of learning visual prompts to adapt pre-trained models to local data distributions and confirm the capability of our proposed approach to address data heterogeneity among multiple clients.



**(a) Office-Caltech10**

**(b) DomainNet**

Figure 5. Comparisons with parameter-efficient fine-tuning methods on (a) Office-Caltech10 and (b) DomainNet datasets.

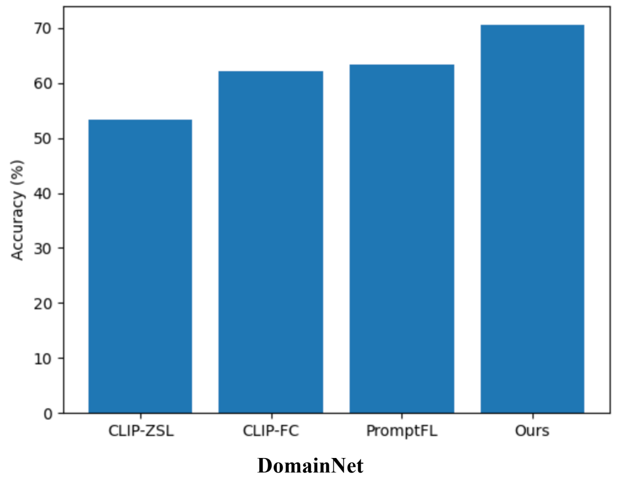


Figure 6. Comparisons with CLIP-based methods on DomainNet dataset. Note that, CLIP-ZSL and CLIP-FC serve as baselines, which are deployed at each client independently without additional prompt learning. The former directly applies the pre-trained CLIP [16] model for inference, while the latter trains a classifier on top of the CLIP visual encoder. In addition, PromtFL [7] updates prompts at each client and then constructs a shared set of global prompts at the server by averaging local prompts.

### B.5. Additional comparisons with CLIP-based methods

In addition to the quantitative comparisons with methods on top of visual foundation models (*e.g.,* supervised ViT [5], MAE [8], and MoCo v3 [2], etc.), we also evaluate it against methods based on CLIP [16], which trained on massive image-text pairs. CLIP-zero denotes directly applying the pre-trained CLIP model to each client without any fine-tuning, while CLIP-FC indicates training fully-connected layers as the classifier locally on top of the frozen CLIP visual encoder. In addition, PromptFL [7] follows CoOp [19] that learns to insert prompts to the text encoder of a frozen CLIP [16] at each client and then averages the prompts trained from local clients at the server. Fig. 6 presents the results on DomainNet, which show that our pFedPG would be preferable among the methods considered. The above results further verify the effectiveness of our personalized prompt generation mechanism for enabling model personalization and handling heterogeneous FL scenarios.

## References

[1] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *NeurIPS*, 2020. 3

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 4

[3] Zhen Chen, Meilu Zhu, Chen Yang, and Yixuan Yuan. Personalized retrogress-resilient framework for real-world medical federated learning. In *MICCAI*, 2021. 1, 2

[4] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *ISBI*, 2018. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3, 4

[6] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical report, 2007. 1

[7] Tao Guo, Song Guo, Junxiao Wang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *arXiv preprint arXiv:2208.11625*, 2022. 4

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3, 4

[9] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3

[10] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fed{bn}: Federated learning on non-{iid} features via local batch normalization. In *ICLR*, 2021. 1

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 3

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017. 3

[13] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 1

[14] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020. 3

[15] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. Rethinking architecture design for tackling data heterogeneity in federated learning. In *CVPR*, 2022. 1, 2, 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[17] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 1

[18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 2018. 1

[19] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 4