Appendix

A. Related Work

Previous studies have concentrated on enhancing the performance of SFL through various optimization techniques. This section reviews the related work on model splitting and resource configuration in heterogeneous edge computing environments.

Model splitting can significantly reduce the workload for clients, primarily focusing on two aspects. First, the model can be split along its depth and width. Some studies reduce the global model from two dimensions based on the client's resources, thereby allocating different client-side models to devices, which decreases the client-side workload [1], [2]. However, splitting the model along its width may reduce its ability to capture fundamental features [3]. Second, the model can be split along its depth. Thapa et al. [4] were the first to design an SFL framework. Based on this research, some works have developed personalized cut layer by considering device computing capabilities. In [5]–[7], an adaptive control method for model splitting was adopted based on a dynamic environment to enhance the performance of SFL. Some studies not only consider the heterogeneity of device computational resources but also take into account the impact of the communication environment [8], [9]. Yu et al. [10] employed reinforcement learning algorithms to address resource configuration and cut layer selection, and designed an semantic-aware auto-encoder to reduces the dimensionality of transmitted data in U-shaped SFL. Conversely, these methods have neither considered leveraging limited server computational resources to accelerate SFL nor addressed the impact of improper splitting schemes on model accuracy.

To reduce training latency, a reasonable resource configuration strategy on the server can significantly accelerate the model training process. Some studies have focused on optimizing SFL from the perspective of server resource configuration. Zhu et al. [11] considered the client-side workload and server resource configuration, dynamically adjusting the cut layer of the client-side model based on the server resource configuration. Khan et al. [12] and Han et al. [13] considered the impact of server resource configuration on SFL costs but overlooked the heterogeneity of devices. Nevertheless, the works fail to consider the scenario where devices with sufficient computational resources do not need to participate in model splitting. Huang et al. [14] comprehensively considered the issues of model offloading and resource configuration, but the offloading decisions were made on the client-side. When a client offloads a significant part of its computational tasks to the server, the server may become inadequate to meet the demand.

Batch size for each client also affects model training efficiency [15]. The batch size affects the overall time to achieve the same performance. Some studies [16]–[18] have analyzed the effect of different batch sizes for various devices to balance their training time. If the batch size is too small, training on the entire dataset requires multiple iterations, thereby increasing the training overhead. As the batch size increases, although the speed of processing the same amount of data improves, the number of epochs needed to achieve performance comparable to that of smaller batch sizes significantly increases. At the same time, it may exceed the capacity of the device, leading to training failure. Additionally, excessively large batch sizes may cause the model to converge to a local optimum, while excessively small batch sizes can also lead to similar issues. Therefore, it is crucial to choose an appropriate batch size based on the size of the local data pool.

B. The Pseudocode of HSFL Workflow

Section II describes the update rules for both the server-side and client-side models, defined respectively as:

$$\mathbf{w}_{s,i}^{h,k+1,l_i} = \mathbf{w}_{s,i}^{h,k,l_i} - \eta \nabla F_{s,i}(\mathbf{w}_{s,i}^{h,k,l_i}), \tag{1}$$

and

$$\mathbf{w}_{c,i}^{h,k+1,l_i} = \mathbf{w}_{c,i}^{h,k,l_i} - \eta \nabla F_{c,i}(\mathbf{w}_{c,i}^{h,k,l_i}). \tag{2}$$

Following the description in Section II, we formalize the workflow as pseudocode, shown in Algorithm 1.

Algorithm 1 The HSFL training framework.

Input: $\mathcal{D}, N, H, \eta$.
Output: $\mathbf{w}^*$.

Clients upload resource profile to register the training task
Initialize the set of c_i^s, global model $\mathbf{w}$, alternative iterate until the set of c_i^s, l_i and b_i are unchanged
for $h = 1$ to H **do**
  for $k = 1$ to τ_i **do**
    for all device $i \in N$ in parallel **do**
      if $l_i = L$ **then**
        FP and BP with the FL training workflow
      else
        / Runs on client **/**
        FP on $\mathbf{w}_{c,i}^{h,k,l_i}$ and generate the $\mathbf{A}_{i,l}$
        Send $\mathbf{A}_{i,l}, \mathbf{Y}_i$ to the server
        / Runs on server **/**
        FP with $\mathbf{A}_{i,l}$ on $\mathbf{w}_{s,i}^{h,k,l_i}$, get the predicted label $\hat{\mathbf{Y}}_i$, and compute loss with $\mathbf{Y}_i$ and $\hat{\mathbf{Y}}_i$
        BP and calculate $\nabla F_{s,i}$
        Update $\mathbf{w}_{s,i}^{h,k+1,l_i}$ by (1)
        Send $\nabla F_{s,i}$ to client i
        / Runs on client **/**
        BP and calculate $\nabla F_{c,i}$ with $\nabla F_{s,i}$
        Update $\mathbf{w}_{c,i}^{h,k+1,l_i}$ by (2)
      end if
    end for
  end for
  Each client sends its local model $\mathbf{w}_{c,i}^{h,l_i}$ to the server
  if server has received all client-side models **then**
    Aggregate all the models to a model $\mathbf{w}^{h+1}$ by (**??**) and distributes the updated model to the clients
  end if
end for

C. The Proof of Theorem 2.1

Proof: Let $\widetilde{\tau} = \max_i \tau_i$. Due to the $\tilde{L}$-Lipschitz smoothness property in Assumption 1, taking the expectation of $f(\mathbf{w}^{h+1})$ in the h-th round communication yields:

$$E^h\left[f(\mathbf{w}^{h+1})\right]$$

$$
\begin{aligned}
&\leq f(\mathbf{w}^h) + \left\langle \nabla f(\mathbf{w}^h), E^h\left[\mathbf{w}^{h+1} - \mathbf{w}^h\right]\right\rangle \\
&\quad + \frac{\tilde{L}}{2} E^h\left[\left\|\mathbf{w}^{h+1} - \mathbf{w}^h\right\|^2\right] \\
&= f(\mathbf{w}^h) + \frac{\tilde{L}}{2} E^h\left[\left\|-\eta \sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k})\right\|^2\right] \\
&\quad + \langle \nabla f(\mathbf{w}^h), E^h\big[-\eta \sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) - \eta\widetilde{\tau}\nabla f(\mathbf{w}^h) \\
&\quad + \eta\widetilde{\tau}\nabla f(\mathbf{w}^h)\big]\rangle \\
&= f(\mathbf{w}^h) - \eta\widetilde{\tau}\left\|\nabla f(\mathbf{w}^h)\right\|^2 \\
&\quad + \underbrace{\left\langle \nabla f(\mathbf{w}^h), E^h\left[-\eta \sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) + \eta\widetilde{\tau}\nabla f(\mathbf{w}^h)\right]\right\rangle}_{A_1} \\
&\quad + \frac{\tilde{L}\eta^2}{2} \underbrace{E^h\left[\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k})\right\|^2\right]}_{A_2}.
\end{aligned} \tag{3}
$$

In (3), the term A_1 is bounded as follows:

$$
\begin{aligned}
&A_1 \\
&= \left\langle \nabla f(\mathbf{w}^h), E^h\left[-\eta \sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) + \eta\widetilde{\tau}\nabla f(\mathbf{w}^h)\right]\right\rangle \\
&= \left\langle \sqrt{\eta\widetilde{\tau}}\nabla f(\mathbf{w}^h), \frac{-\sqrt{\eta}}{\sqrt{\widetilde{\tau}}} E^h\left[\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) - \widetilde{\tau}\nabla f(\mathbf{w}^h)\right]\right\rangle \\
&\overset{(a_1)}{=} \frac{\eta\widetilde{\tau}}{2}\left\|\nabla f(\mathbf{w}^h)\right\|^2 \\
&\quad + \frac{\eta}{2\widetilde{\tau}} \underbrace{E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) - \widetilde{\tau}\nabla f(\mathbf{w}^h)\right\|^2}_{B_1} \\
&\quad - \frac{\eta}{2\widetilde{\tau}} \underbrace{E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k})\right\|^2}_{A_2},
\end{aligned} \tag{4}
$$

where (a_1) is derived from $\langle x, y\rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x-y\|^2$. In (4), the term B_1 can be bounded as follows:

$$
\begin{aligned}
&B_1 \\
&= E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) - \widetilde{\tau}\nabla f(\mathbf{w}^h)\right\|^2 \\
&= E^h\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) \\
&\quad - \sum_{i=1}^{N} p_i(\sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}^h) + \sum_{\tau=\tau_i}^{\widetilde{\tau}-1} \nabla F_i(\mathbf{w}^h))\|^2 \\
&= E^h\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} (\nabla F_i(\mathbf{w}_i^{h,k}) - \nabla F_i(\mathbf{w}^h)) \\
&\quad - \sum_{i=1}^{N} p_i(\widetilde{\tau} - \tau_i)(\nabla F_i(\mathbf{w}^h) - \nabla f(\mathbf{w}^h)) \\
&\quad - \sum_{i=1}^{N} p_i(\widetilde{\tau} - \tau_i)\nabla f(\mathbf{w}^h)\|^2 \\
&\overset{(b_1)}{\leq} 3E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} (\nabla F_i(\mathbf{w}_i^{h,k}) - \nabla F_i(\mathbf{w}^h))\right\|^2 \\
&\quad + 3E^h\left\|\sum_{i=1}^{N} p_i(\widetilde{\tau} - \tau_i)(\nabla F_i(\mathbf{w}^h) - \nabla f(\mathbf{w}^h))\right\|^2 \\
&\quad + 3E^h\left\|\sum_{i=1}^{N} p_i(\widetilde{\tau} - \tau_i)\nabla f(\mathbf{w}^h)\right\|^2 \\
&\overset{(b_2)}{\leq} 3(\sum_{i=1}^{N}\tau_i)\sum_{i=1}^{N} p_i^2 \sum_{\tau=0}^{\tau_i-1} E^h\left\|\nabla F_i(\mathbf{w}_i^{h,k}) - \nabla F_i(\mathbf{w}^h)\right\|^2 \\
&\quad + 3N\sum_{i=1}^{N} E^h\left\|p_i(\widetilde{\tau} - \tau_i)(\nabla F_i(\mathbf{w}^h) - \nabla f(\mathbf{w}^h))\right\|^2 \\
&\quad + 3N(\sum_{i=1}^{N} p_i^2(\widetilde{\tau} - \tau_i)^2)E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2 \\
&\overset{(b_3)}{\leq} 3\sigma_L^2(\sum_{i=1}^{N}\tau_i)(\sum_{i=1}^{N} p_i^2\tau_i) + 3N\sigma_G^2(\sum_{i=1}^{N} p_i^2(\widetilde{\tau} - \tau_i)^2) \\
&\quad + 3N(\sum_{i=1}^{N} p_i^2(\widetilde{\tau} - \tau_i)^2)E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2,
\end{aligned} \tag{5}
$$

where (b_1) and (b_2) are derived from $E[\|x_1 + \cdot + x_n\|^2] \leq nE[\|x_1\|^2 + \cdot + \|x_n\|^2]$, (b_3) is derived from Assumption 3. Substituting B_1 into A_1, we get:

$$
\begin{aligned}
&A_1 \\
&\leq E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2\left(\frac{\eta\widetilde{\tau}}{2} + \frac{3\eta N(\sum_{i=1}^{N} p_i^2(\widetilde{\tau} - \tau_i)^2)}{2\widetilde{\tau}}\right) \\
&\quad + \frac{3\eta\sigma_L^2(\sum_{i=1}^{N}\tau_i)(\sum_{i=1}^{N} p_i^2\tau_i) + 3\eta N\sigma_G^2(\sum_{i=1}^{N} p_i^2(\widetilde{\tau} - \tau_i)^2)}{2\widetilde{\tau}} \\
&\quad - \frac{\eta}{2\widetilde{\tau}} A_2
\end{aligned} \tag{6}
$$

Next, we derive the term A_2 in (3) and (6), where A_2 is bounded as follows:

$$
\begin{aligned}
&A_2 \\
&= E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k})\right\|^2 \\
&= E^h\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} (\nabla F_i(\mathbf{w}_i^{h,k}) - \nabla F_i(\mathbf{w}^h) + \nabla F_i(\mathbf{w}^h) \\
&\quad - \nabla f(\mathbf{w}^h) + \nabla f(\mathbf{w}^h))\|^2 \\
&\leq 3E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}_i^{h,k}) - \nabla F_i(\mathbf{w}^h)\right\|^2 \\
&\quad + 3E^h\left\|\sum_{i=1}^{N} p_i \sum_{\tau=0}^{\tau_i-1} \nabla F_i(\mathbf{w}^h) - \nabla f(\mathbf{w}^h)\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+3E^h\left\|\sum_{i=1}^{N}p_i\sum_{\tau=0}^{\tau_i-1}\nabla f(\mathbf{w}^h)\right\|^2\\
\leq &3(\sigma_L^2+\sigma_G^2)(\sum_{i=1}^{N}\tau_i)(\sum_{i=1}^{N}p_i^2\tau_i)+3N(\sum_{i=1}^{N}p_i^2\tau_i^2)E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2
\end{aligned}
\tag{7}
$$

Substituting A_1 and A_2 into (3), we yields:

$$
\begin{aligned}
&E^h\left[f(\mathbf{w}^{h+1})\right]\\
&\leq f(\mathbf{w}^h)-\frac{\eta}{2}[(\tilde{\tau}-3\tilde{L}\eta N\sum_{i=1}^{N}p_i^2\tau_i^2)\\
&+3N\sum_{i=1}^{N}p_i^2(2\tau_i-\tilde{\tau})]\left\|\nabla f(\mathbf{w}^h)\right\|^2\\
&+\frac{3\eta N\sigma_G^2}{2\tilde{\tau}}(\sum_{i=1}^{N}p_i^2(\tilde{\tau}-\tau_i)^2)+\frac{3\tilde{L}\eta^2}{2}(\sigma_L^2+\sigma_G^2)(\sum_{i=1}^{N}\tau_i)(\sum_{i=1}^{N}p_i^2\tau_i)\\
&\overset{(a_2)}{\leq} f(\mathbf{w}^h)-\frac{\eta}{2}\theta\left\|\nabla f(\mathbf{w}^h)\right\|^2+\frac{\eta\phi}{2}.
\end{aligned}
\tag{8}
$$

If $\tau_i > \frac{\tilde{\tau}}{2}$ and $\eta\tilde{\tau} < \frac{1}{3N\tilde{L}}$ for any client i, it can be concluded that $\tilde{\tau}-3\tilde{L}\eta N\sum_{i=1}^{N}p_i^2\tau_i^2 > 0$ and $\sum_{i=1}^{N}p_i^2(2\tau_i-\tilde{\tau}) > 0$, consequently there exists a constant θ such that $(\tilde{\tau}-3\tilde{L}\eta N\sum_{i=1}^{N}p_i^2\tau_i^2)+3N\sum_{i=1}^{N}p_i^2(2\tau_i-\tilde{\tau}) > \theta > 0$. (a_2) can be derived from applying the above analysis and setting $\phi = \frac{3N\sigma_G^2}{\tilde{\tau}}(\sum_{i=1}^{N}p_i^2(\tilde{\tau}-\tau_i)^2)+3\tilde{L}\eta(\sigma_L^2+\sigma_G^2)(\sum_{i=1}^{N}\tau_i)(\sum_{i=1}^{N}p_i^2\tau_i)$.

Rearranging and summing from $h=0,\cdots,H-1$, we have:

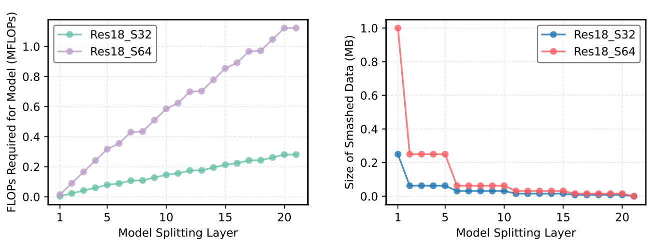
$$
\sum_{h=0}^{H-1}\frac{\eta\theta}{2}E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2 \leq f(\mathbf{w}^0)-f(\mathbf{w}^H)+\frac{\eta H\phi}{2}, \tag{9}
$$

which implies

$$
\min_{h\in\{1,\cdots,H\}} E^h\left\|\nabla f(\mathbf{w}^h)\right\|^2 \leq \frac{2(f_0-f_*)}{\eta\theta H}+\frac{\phi}{\theta}. \tag{10}
$$

This completes the proof. ■

D. The Information of Renset-18 Network



(a) FLOPs Required for Model (b) Size of Smashed Data

Fig. 1: Functions changes of the cut layer.

Fig. 1 presents the size of smashed data and Floating Point Operations (FLOPs) of Resnet-18 with different input size. "Res18_32" indicates that the ResNet-18 model is employed, and the input images are of size 3×32×32 pixels with a batch size of 1. Similarly, "Res18_64" represents an input image size of 3×64×64 pixels.

References

[1] E. Diao, J. Ding, and V. Tarokh, "Heterofl: Computation and communication efficient federated learning for heterogeneous clients," *arXiv preprint arXiv:2010.01264*, 2020.

[2] F. Ilhan, G. Su, and L. Liu, "Scalefl: Resource-adaptive federated learning with heterogeneous clients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24532–24541, 2023.

[3] M. Kim, S. Yu, S. Kim, and S.-M. Moon, "Depthfl: Depthwise federated learning for heterogeneous clients," in *The Eleventh International Conference on Learning Representations*, 2023.

[4] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When federated learning meets split learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8485–8493, 2022.

[5] Z. Lin, G. Qu, W. Wei, X. Chen, and K. K. Leung, "Adaptsfl: Adaptive split federated learning in resource-constrained edge networks," *arXiv preprint arXiv:2403.13101*, 2024.

[6] T. Xia, Y. Deng, S. Yue, J. He, J. Ren, and Y. Zhang, "Hsfl: an efficient split federated learning framework via hierarchical organization," in *2022 18th International Conference on Network and Service Management (CNSM)*, pp. 1–9, IEEE, 2022.

[7] D. Yan, M. Hu, Z. Xia, Y. Yang, J. Xia, X. Xie, and M. Chen, "Have your cake and eat it too: Toward efficient and accurate split federated learning," *arXiv preprint arXiv:2311.13163*, 2023.

[8] C. Xu, J. Li, Y. Liu, Y. Ling, and M. Wen, "Accelerating split federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 23, no. 6, pp. 5587–5599, 2024.

[9] H. Ao, H. Tian, and W. Ni, "Federated split learning for edge intelligence in resource-constrained wireless networks," *IEEE Transactions on Consumer Electronics*, 2024.

[10] L. Yu, Z. Chang, Y. Jia, and G. Min, "Model partition and resource allocation for split learning in vehicular edge networks," *arXiv preprint arXiv:2411.06773*, 2024.

[11] G. Zhu, Y. Deng, X. Chen, H. Zhang, Y. Fang, and T. F. Wong, "Esfl: Efficient split federated learning over resource-constrained heterogeneous wireless devices," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27153–27166, 2024.

[12] L. U. Khan, M. Guizani, A. Al-Fuqaha, C. S. Hong, D. Niyato, and Z. Han, "A joint communication and learning framework for hierarchical split federated learning," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 268–282, 2024.

[13] D.-J. Han, D.-Y. Kim, M. Choi, D. Nickel, J. Moon, M. Chiang, and C. G. Brinton, "Federated split learning with joint personalization-generalization for inference-stage optimization in wireless edge networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 7048–7065, 2024.

[14] B. Huang, H. Zhao, L. Wang, W. Qian, Y. Yin, and S. Deng, "Decentralized proactive model offloading and resource allocation for split and federated learning," *IEEE Internet of Things Journal*, 2024.

[15] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017.

[16] Y. Liao, Y. Xu, H. Xu, Z. Yao, L. Wang, and C. Qiao, "Accelerating federated learning with data and model parallelism in edge computing," *IEEE/ACM Transactions on Networking*, vol. 32, no. 1, pp. 904–918, 2024.

[17] Y. Liao, Y. Xu, H. Xu, L. Wang, Z. Yao, and C. Qiao, "Mergesfl: Split federated learning with feature merging and batch size regulation," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 2054–2067, IEEE, 2024.

[18] Z. Ma, Y. Xu, H. Xu, Z. Meng, L. Huang, and Y. Xue, "Adaptive batch size for federated learning in resource-constrained edge computing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 37–53, 2021.