A. Additional Dataset Information

Client	1	2	3	4	5	6	Global
Train	10	16	18	18	25	50	137
Val	5	8	9	9	12	25	68
Test	4	8	8	8	13	24	65

Table 8. Prostate dataset: number of data (3D image) in each client.

Figure 5. Representative original 2D image in retinal dataset (low data similarity). First row: client 1 to 3. Second row: client 4 to 6.

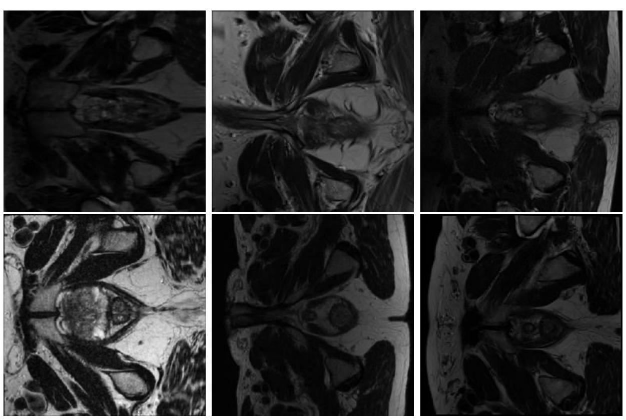

Figure 6. Representative original 2D image slices in prostate dataset (high data similarity). First row: client 1 to 3. Second row: client 4 to 6. E.g., the first slice comes from a 3D image in client 1.

Algorithm 3 FedSM-extra training.

Input: local dataset $\mathcal{D}_k$, rounds R, number of sites K, learning rate η, coefficient λ, client weight $\frac{n_k}{n}$.
Initialize: global model w_g, personalized model $w_{p,k}$, model selector w_s, base optimizer $\text{OPT}(\cdot)$
for round $r = 1, 2, \cdots, R$ **do**
  SERVER: send models $(w_g, w_{p,k})$ to client k.
  for CLIENT $k \in \{1, 2, \cdots, K\}$ in parallel **do**
    initialize $w_{g,k} \leftarrow w_g$
    for batch $(x, y) \in \mathcal{D}_n$ **do**
      $w_{g,k} \leftarrow \text{OPT}(w_{g,k}, \eta, \nabla_{w_{g,k}} L(f(w_{g,k}; x), y))$
      $w_{p,k} \leftarrow \text{OPT}(w_{p,k}, \eta, \nabla_{w_{p,k}} L(f(w_{p,k}; x), y))$
    end for
    send $(w_{g,k}, w_{p,k})$ to server
  end for
  SERVER: $w_g \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{g,k}$ *// FedAvg*
  SERVER: $\forall k \in \{1, 2, \cdots, K\}$, $w_{p,k} \leftarrow \lambda w_{p,k} + (1-\lambda)\frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K} w_{p,k'}$ *// SoftPull*
end for
// extra training rounds
SERVER: send models $(w_g, w_{p,n})$ to clients.
for round $r = 1, 2, \cdots, \Delta R$ **do**
  SERVER: send model w_s to clients.
  for CLIENT $k \in \{1, 2, \cdots, K\}$ in parallel **do**
    Initialize $w_{s,k} \leftarrow w_s$
    for batch $(x, y) \in \mathcal{D}_n$ **do**
      // y_s from Eq. (1)
      $w_{s,k} \leftarrow \text{OPT}(w_{s,k}, \eta_s, \nabla_{w_{s,k}} L_s(f_s(w_{s,k}; x), y_s))$
    end for
    send $w_{s,k}$ to server
  end for
  SERVER: $w_s \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{s,k}$
end for
Output: model $(w_g, \{w_{p,k}\}_{k=1}^{K}, w_s)$

Algorithm 4 FedSM-extra inference.

Input: data x, model $(w_g, \{w_{p,k}\}_{k=1}^{K}, w_s)$
$\widehat{y}_s = f_s(w_s; x)$
$k = \arg\max(\widehat{y}_s) \in \{0, 1, \cdots, K\}$
if $k > 0$ **then**
  $\widehat{y} = f(w_{p,k}; x)$
else
  $\widehat{y} = f(w_g; x)$
end if
Output: $\widehat{y}$

B. FedSM-extra Algorithm

For the training of FedSM-extra, we train the global model and personalized models first, and then train the model selector, which incurs extra ΔR training rounds. In each training round of FedSM-extra, the communication cost is $2w_g$ in the previous R rounds (the global and personalized models have the same model architecture). It becomes w_s in the extra ΔR training rounds.

For the inference of FedSM-extra, both the global model and personalized models can be selected. Therefore $k \in$

$\{0, 1, \cdots, K\}$ (For FedSM, $k \in \{1, 2, \cdots, K\}$).

C. Proof of SoftPull Convergence

Let the current/total training rounds be r/R, current/total local training steps be m/M, the current/total global training step be t/T. We denote the personalized model during training as $w_{p,k}^{r,m}$. For simplicity we use f_k to denote loss $L_{\mathcal{D}_k}$.

After the local training in the last training round $r-1$ finishes, we get model $w_{p,k}^{r-1,M}$ and want to

$$\min \sum_{k=1}^{K} f_k\left(\frac{1}{\lambda} w_{p,k}^{r-1,M} - \frac{1-\lambda}{\lambda} \frac{1}{K-1} \sum_{k'=1,k'\neq k}^{K} w_{p,k'}^{r-1,M}\right) \tag{15}$$

In the beginning of the current training round r, from Eq. (11), we will have

$$w_{p,k}^{r,0} = \lambda w_{p,k}^{r-1,M} + (1-\lambda)\frac{1}{K-1} \sum_{k'=1,k'\neq k}^{K} w_{p,k'}^{r-1,M} \tag{16}$$

In the current training round r, we consider two stages. The first stage is a transition from then end of training round $r-1$ to the start of the current training round r, while the second stage is the start to the end of current training round r. Let $\lambda' = \frac{K\lambda - 1}{K-1}$, $1-\lambda' = \frac{K}{K-1}(1-\lambda)$, then

$$w_{p,k}^{r,0} = \lambda' w_{p,k}^{r-1,M} + (1-\lambda')\overline{w}_{p,k}^{r-1,M} \tag{17}$$

where the bar denotes an average over all clients $k \in \{1, 2, \cdots, K\}$. It can be clearly seen that when the data distributions of clients are very similar, we set $\lambda = \frac{1}{K}$, $\lambda' = 0$, i.e., the "hard averaging" in FedAvg. When the data distributions are not similar at all, we set $\lambda = 1$, $\lambda' = 1$ to only do local training. In other circumstances, theoretically we should set $\lambda \in [\frac{1}{K}, 1]$, $\lambda' \in [0, 1]$ according to the data similarity.

C.1. Difference

Suppose the stochastic gradient at iteration (r, m) is $\nabla f_k(w_{p,k}^{r,m}, x_{p,k}^{r,m})$ and the expected gradient is $\nabla f_k(w_{p,k}^{r,m}) = \mathbb{E}_{x_{p,k}^{r,m} \in \mathcal{D}_k} \nabla f_k(w_{p,k}^{r,m}, x_{p,k}^{r,m}) = \mathbb{E}\nabla f_k(w_{p,k}^{r,m}, x_{p,k}^{r,m})$. We need to bound

$$\begin{aligned}
&\|(w_{p,k}^{r+1,0} - w_{p,k}^{r,M})\|_2^2 \\
&= \|(1-\lambda)(w_{p,k}^{r,M} - \frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K} w_{p,k'}^{r,M})\|_2^2 \\
&= (1-\lambda)^2 \|w_{p,k}^{r,M} - \frac{1}{K-1}(K\overline{w}_{p,k}^{r,M} - w_{p,k}^{r,M})\|_2^2 \\
&= \frac{(1-\lambda)^2 K^2}{(K-1)^2}\|w_{p,k}^{r,M} - \overline{w}_{p,k}^{r,M}\|_2^2
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
&\mathbb{E}\|w_{p,k}^{r,M} - \overline{w}_{p,k}^{r,M}\|_2^2 \\
&= \eta^2 \mathbb{E}\| \sum_{r'=0}^{r} (\lambda')^{r-r'} \sum_{m=0}^{M-1} [\nabla f_k(w_{p,k}^{r',m}, x_{p,k}^{r',m}) \\
&\quad - \overline{\nabla f_k}(w_{p,k}^{r',m}, x_{p,k}^{r',m})]\|_2^2 \\
&\leq \eta^2 (\sum_{r'=0}^{r} (\lambda')^{r-r'})^2 \mathbb{E}\| \sum_{r'=0}^{r} \frac{(\lambda')^{r-r'}}{\sum_{r'=0}^{r}(\lambda')^{r-r'}} \sum_{m=0}^{M-1} \\
&\quad [\nabla f_k(w_{p,k}^{r',m}, x_{p,k}^{r',m}) - \overline{\nabla f_k}(w_{p,k}^{r',m}, x_{p,k}^{r',m})]\|_2^2 \\
&\leq \eta^2 (\sum_{r'=0}^{r} (\lambda')^{r-r'}) \sum_{r'=0}^{r} (\lambda')^{r-r'} \mathbb{E}\| \sum_{m=0}^{M-1} \\
&\quad [\nabla f_k(w_{p,k}^{r',m}, x_{p,k}^{r',m}) - \overline{\nabla f_k}(w_{p,k}^{r',m}, x_{p,k}^{r',m})]\|_2^2 \\
&\leq M\eta^2 (\sum_{r'=0}^{r} (\lambda')^{r-r'}) \sum_{r'=0}^{r} (\lambda')^{r-r'} \sum_{m=0}^{M-1} \\
&\quad \mathbb{E}\|\nabla f_k(w_{p,k}^{r',m}, x_{p,k}^{r',m}) - \overline{\nabla f_k}(w_{p,k}^{r',m}, x_{p,k}^{r',m})\|_2^2 \\
&\leq 2M^2(G^2+\sigma^2)\eta^2 (\sum_{r'=0}^{r} (\lambda')^{r-r'})^2 \\
&\leq 2M^2(G^2+\sigma^2)\eta^2 [\frac{1-(\lambda')^{r+1}}{1-\lambda'}]^2
\end{aligned} \tag{19}$$

where $\mathbb{E}\|\nabla f_k(w_{p,k}^{r',m}, x_{p,k}^{r',m})\|_2^2 \leq 2(G^2+\sigma^2)$ based on Assumptions 3 and 2. Then

$$\begin{aligned}
&\mathbb{E}\|(w_{p,k}^{r+1,0} - w_{p,k}^{r,M})\|_2^2 \\
&\leq \frac{[1-(\lambda')^{r+1}]^2(1-\lambda)^2 K^2}{(1-\lambda')^2(K-1)^2} 2M^2(G^2+\sigma^2)\eta^2 \\
&= [1-(\lambda')^{r+1}]^2 2M^2(G^2+\sigma^2)\eta^2
\end{aligned} \tag{20}$$

C.2. Local Objective

Here we consider the local objective function to optimize. From $(r, 0)$ to (r, M), i.e. $m \in \{0, 1, \cdots, M-1\}$, due to the Lipschitz smooth assumption we have

$$\begin{aligned}
&f_k(w_{k,p}^{r,m+1}) - f_k(w_{k,p}^{r,m}) \\
&\leq \langle \nabla f_k(w_{k,p}^{r,m}), w_{k,p}^{r,m+1} - w_{k,p}^{r,m}\rangle + \frac{L}{2}\|w_{k,p}^{r,m+1} - w_{k,p}^{r,m}\|_2^2 \\
&= -\eta\langle \nabla f_k(w_{k,p}^{r,m}), \nabla f_k(w_{k,p}^{r,m}, x_{k,p}^{r,m})\rangle \\
&\quad + \frac{\eta^2 L}{2}\|\nabla f_k(w_{k,p}^{r,m}, x_{k,p}^{r,m})\|_2^2 \\
&= -\eta\langle \nabla f_k(w_{k,p}^{r,m}), \nabla f_k(w_{k,p}^{r,m}, x_{k,p}^{r,m})\rangle \\
&\quad + \frac{\eta^2 L}{2}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 + \frac{\eta^2 L\sigma^2}{2}
\end{aligned} \tag{21}$$

Take the expectation and suppose $\eta \leq \frac{1}{L}$,

$$\begin{aligned}
&\mathbb{E}[f_k(w_{k,p}^{r,m+1}) - f_k(w_{k,p}^{r,m})] \\
&\leq -\eta(1-\frac{\eta L}{2})\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 + \frac{\eta^2 L\sigma^2}{2} \\
&\leq -\frac{\eta}{2}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 + \frac{\eta^2 L\sigma^2}{2}
\end{aligned} \tag{22}$$

$$\begin{aligned}
&\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\leq \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,m}) - f_k(w_{k,p}^{r,m+1})] + \eta L\sigma^2
\end{aligned} \tag{23}$$

$$\begin{aligned}
&\sum_{m=0}^{M-1}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\leq \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r,M})] + M\eta L\sigma^2
\end{aligned} \tag{24}$$

While from (r, M) to $(r+1, 0)$, we have

$$\begin{aligned}
&f_k(w_{k,p}^{r+1,0}) - f_k(w_{k,p}^{r,M}) \\
&\leq \langle \nabla f_k(w_{k,p}^{r,M}), w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\rangle + \frac{L}{2}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
&\leq \frac{\eta}{8}\|\nabla f_k(w_{k,p}^{r,M})\|_2^2 + (\frac{2}{\eta} + \frac{L}{2})\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
&\leq \frac{\eta}{4}\|\nabla f_k(w_{k,p}^{r,M-1})\|_2^2 + \frac{\eta L^2}{4}\|w_{k,p}^{r,M} - w_{k,p}^{r,M-1}\|_2^2 \\
&\quad + (\frac{2}{\eta} + \frac{L}{2})\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
&= \frac{\eta}{4}\|\nabla f_k(w_{k,p}^{r,M-1})\|_2^2 + \frac{\eta^3 L^2}{4}\|\nabla f_k(w_{k,p}^{r,M-1}, x_{k,p}^{r,M-1})\|_2^2 \\
&\quad + (\frac{2}{\eta} + \frac{L}{2})\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2
\end{aligned} \tag{25}$$

Therefore, from $(r, 0)$ to $(r+1, 0)$, we have

$$\begin{aligned}
&\sum_{m=0}^{M-1}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\leq \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r+1,0})] + M\eta L\sigma^2 \\
&\quad + \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r+1,0}) - f_k(w_{k,p}^{r,M})] \\
&\leq \frac{2}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r+1,0})] + M\eta L\sigma^2 \\
&\quad + \frac{1}{2}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,M-1})\|^2 + \eta^2 L^2(G^2+\sigma^2) \\
&\quad + (\frac{4}{\eta^2} + \frac{L}{\eta})\mathbb{E}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2
\end{aligned} \tag{26}$$

$$\begin{aligned}
&\sum_{m=0}^{M-1}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\leq \frac{4}{\eta}\mathbb{E}[f_k(w_{k,p}^{r,0}) - f_k(w_{k,p}^{r+1,0})] + 2M\eta L\sigma^2 \\
&\quad + 2\eta^2 L^2(G^2+\sigma^2) + (\frac{8}{\eta^2} + \frac{2L}{\eta})\mathbb{E}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2
\end{aligned} \tag{27}$$

From $r = 0$ to $R-1$,

$$\begin{aligned}
&\frac{1}{RM}\sum_{r=0}^{R-1}\sum_{m=0}^{M-1}\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\leq \frac{4\mathbb{E}[f_k(w_{k,p}^{0,0}) - f_k(w_{k,p}^{R,0})]}{\eta RM} + 2\eta L\sigma^2 \\
&\quad + \frac{2\eta^2 L^2(G^2+\sigma^2)}{M} + \frac{1}{RM}(\frac{8}{\eta^2} + \frac{2L}{\eta}) \\
&\quad \cdot \sum_{r=0}^{R-1}\mathbb{E}\|w_{k,p}^{r+1,0} - w_{k,p}^{r,M}\|_2^2 \\
&= \frac{4\mathbb{E}[f_k(w_{k,p}^{0,0}) - f_k(w_{k,p}^{R,0})]}{\eta RM} + 2\eta L\sigma^2 \\
&\quad + \frac{2\eta^2 L^2(G^2+\sigma^2)}{M} + \frac{1}{RM}(\frac{8}{\eta^2} + \frac{2L}{\eta}) \\
&\quad \cdot \frac{(1-\lambda)^2 K^2}{(K-1)^2}\sum_{r=0}^{R-1}\mathbb{E}\|w_{k,p}^{r,M} - \overline{w}_{k,p}^{r,M}\|_2^2
\end{aligned} \tag{28}$$

C.3. Proposed Objective

Here we consider our proposed personalized FL objective function to optimize. For simplicity of notation, let

$$u_k^{r,m} = \frac{1}{\lambda}w_{p,k}^{r,m} - \frac{1-\lambda}{\lambda}\frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K} w_{p,k'}^{r,m} \tag{29}$$

Then

$$\begin{aligned}
&u_k^{r,m} - w_{p,k}^{r,m} = \frac{1-\lambda}{\lambda}(w_{p,k}^{r,m} - \frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K} w_{p,k'}^{r,m}) \\
&= \frac{1-\lambda}{\lambda}\frac{K}{K-1}(w_{p,k}^{r,m} - \overline{w}_{p,k}^{r,m})
\end{aligned} \tag{30}$$

Now we bound the gradient of the proposed objective.

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla_{w_{p,k}^{r,m}}\sum_{k'=1}^{K}f_{k'}(u_{k'}^{r,m})\|_2^2 \\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{\lambda}\nabla f_k(u_k^{r,m})-\frac{1-\lambda}{\lambda}\frac{1}{K-1}\nabla f_{k'}(u_{k'}^{r,m})\|_2^2 \\
&\leq\frac{2}{K}\sum_{k=1}^{K}(\frac{1}{\lambda^2}+\frac{(1-\lambda)^2}{\lambda^2(K-1)})\mathbb{E}\|\nabla f_k(u_k^{r,m})\|_2^2 \\
&=\frac{2}{K}\sum_{k=1}^{K}(\frac{1}{\lambda^2}+\frac{(1-\lambda)^2}{\lambda^2(K-1)})[\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\quad+L^2\mathbb{E}\|u_k^{r,m}-w_{k,p}^{r,m}\|_2^2] \\
&=(\frac{1}{\lambda^2}+\frac{(1-\lambda)^2}{\lambda^2(K-1)})\frac{2}{K}\sum_{k=1}^{K}[\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2 \\
&\quad+\frac{L^2(1-\lambda)^2K^2}{\lambda^2(K-1)^2}\mathbb{E}\|w_{k,p}^{r,m}-\overline{w}_{k,p}^{r,m}\|_2^2]
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
&\frac{1}{KRM}\sum_{r=0}^{R-1}\sum_{m=0}^{M-1}\sum_{k=1}^{K}\mathbb{E}\|\nabla_{w_{p,k}^{r,m}}\sum_{k'=1}^{K}f_{k'}(u_{k'}^{r,m})\|_2^2 \\
&\leq(\frac{1}{\lambda^2}+\frac{(1-\lambda)^2}{\lambda^2(K-1)})\frac{2}{KRM}\sum_{k=1}^{K}\sum_{r=0}^{R-1}\sum_{m=0}^{M-1} \\
&[\mathbb{E}\|\nabla f_k(w_{k,p}^{r,m})\|_2^2+\frac{L^2(1-\lambda)^2K^2}{(K-1)^2}\mathbb{E}\|w_{k,p}^{r,m}-\overline{w}_{k,p}^{r,m}\|_2^2] \\
&\leq 2(\frac{1}{\lambda^2}+\frac{(1-\lambda)^2}{\lambda^2(K-1)})[\frac{\frac{4}{K}\sum_{k=1}^{K}(f_k^0-f_k^*)}{\eta RM} \\
&\quad+2\eta L\sigma^2+\frac{2\eta^2L^2(G^2+\sigma^2)}{M} \\
&\quad+\frac{1}{KRM}(\frac{8}{\eta^2}+\frac{2L}{\eta})\frac{(1-\lambda)^2K^2}{(K-1)^2} \\
&\quad\cdot\sum_{k=1}^{K}\sum_{r=0}^{R-1}\mathbb{E}\|w_{k,p}^{r,M}-\overline{w}_{k,p}^{r,M}\|_2^2 \\
&\quad+\frac{1}{KRM}\frac{L^2(1-\lambda)^2K^2}{\lambda^2(K-1)^2}\sum_{k=1}^{K}\sum_{r=0}^{R-1}\sum_{m=0}^{M-1}\mathbb{E}\|w_{k,p}^{r,m}-\overline{w}_{k,p}^{r,m}\|_2^2]
\end{aligned}
\tag{32}
$$

which converges to

$$
\begin{aligned}
&\mathcal{O}(\frac{1}{\eta RM\lambda^2}+\frac{(1-\lambda)^2}{KRM\eta^2\lambda^2}\sum_{K=1}^{K}\sum_{r=0}^{R-1}\mathbb{E}\|w_{k,p}^{r,M}-\overline{w}_{k,p}^{r,M}\|_2^2 \\
&\quad+\frac{(1-\lambda)^2}{KRM\lambda^4}\sum_{k=1}^{K}\sum_{r=0}^{R-1}\sum_{m=0}^{M-1}\mathbb{E}\|w_{k,p}^{r,m}-\overline{w}_{k,p}^{r,m}\|_2^2) \\
&=\mathcal{O}(\frac{1}{\eta RM\lambda^2}+\frac{M\sum_{r=0}^{R-1}(1-\lambda)^2}{R\lambda^2} \\
&\quad+\frac{M^2\eta^2\sum_{r=0}^{R-1}(1-\lambda)^2}{R\lambda^4})
\end{aligned}
\tag{33}
$$

Suppose $\eta=\mathcal{O}(\frac{1}{\sqrt{RM}})$ and $M=\mathcal{O}(R^{\frac{1}{3}})$, the convergence rate is $\mathcal{O}(\frac{1}{\lambda^4\sqrt{RM}})$ with an error $\mathcal{O}(\frac{M\sum_{r=0}^{R-1}(1-\lambda)^2}{R\lambda^2})$.

D. Additional Experimental Results

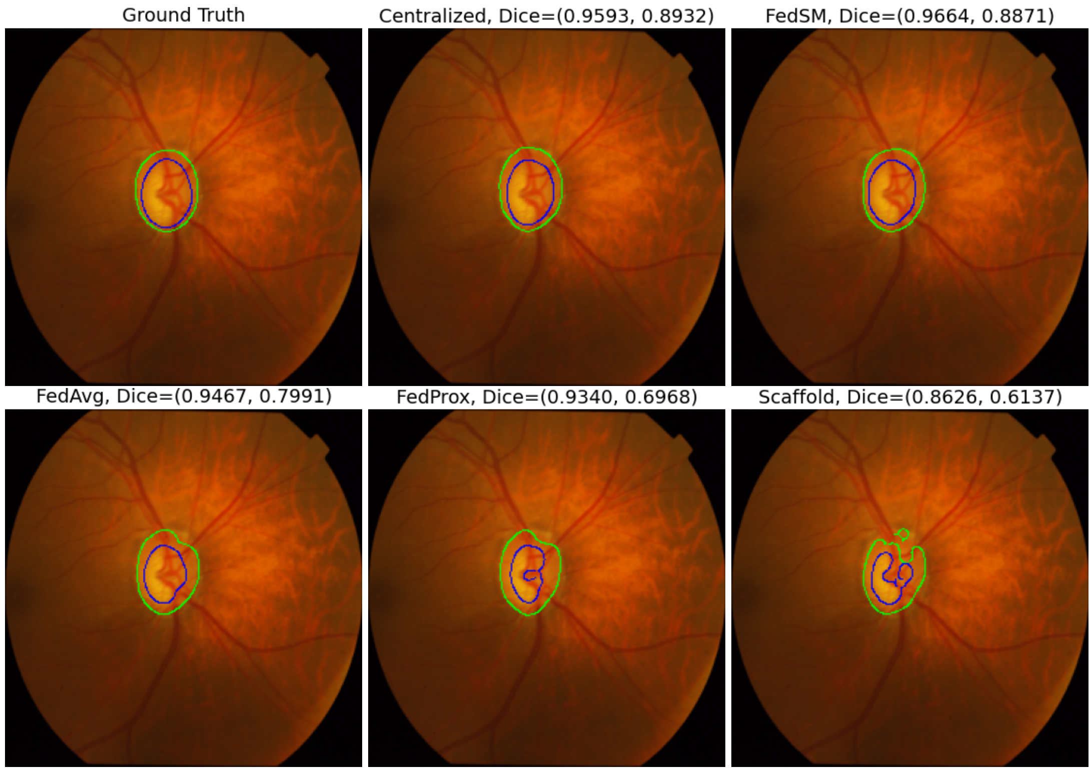

Figure 7. Visual comparison of retinal disc (green) and cup (blue) segmentation. Dice denotes the retinal disc and cup Dice coefficient.

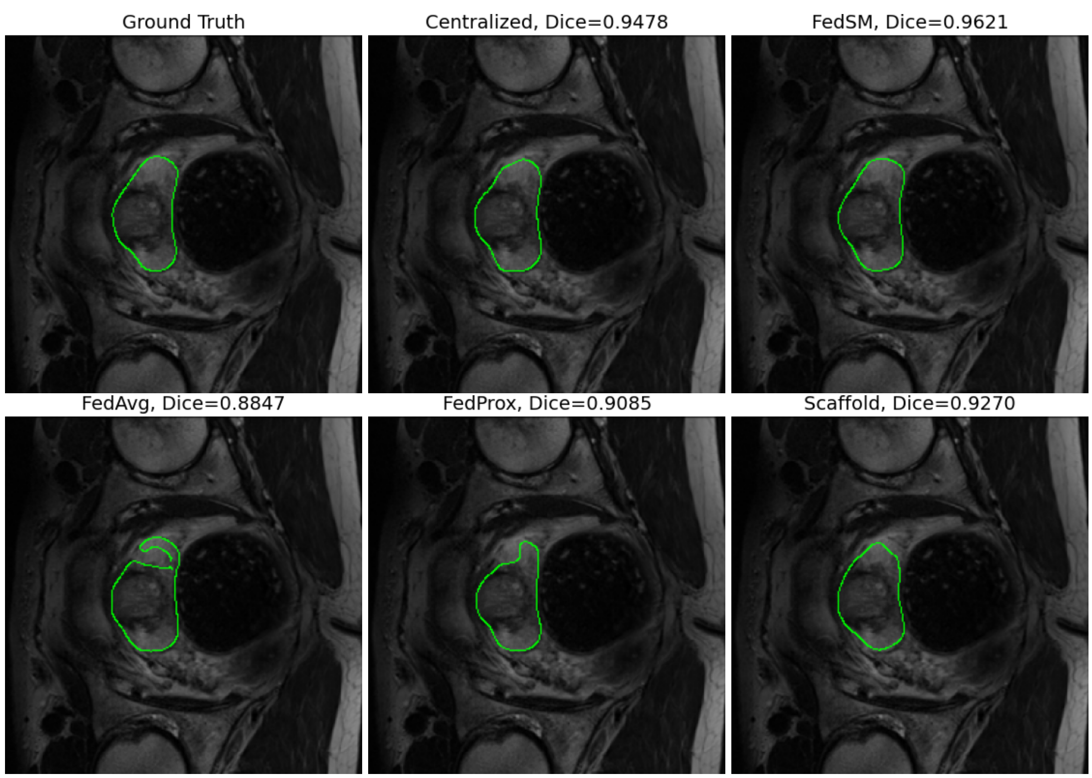

Figure 8. Visual comparison of prostate (green) segmentation. Dice denotes the Dice coefficient.

Method	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Client Avg Dice	Global Dice
Centralized	0.9628	0.9486	0.9489	0.9539	0.9242	0.9565	0.9492	0.9522
Client 1 Local	0.9454	0.4357	0.8956	0.6073	0.4464	0.8409	0.6952	0.7129
Client 2 Local	0.2936	0.9431	0.1099	0.8371	0.2439	0.4575	0.4809	0.5589
Client 3 Local	0.9420	0.3998	0.9468	0.6399	0.4349	0.8256	0.6982	0.7120
Client 4 Local	0.6830	0.9400	0.4805	0.9526	0.3088	0.7803	0.6909	0.7796
Client 5 Local	0.6102	0.2169	0.4601	0.2518	0.9033	0.7064	0.5248	0.5373
Client 6 Local	0.8806	0.7937	0.8354	0.8475	0.4413	0.9547	0.7922	0.8555
FedAvg	0.9554	0.9410	0.9372	0.9535	0.8653	0.9549	0.9346	0.9444
FedProx	0.9447	0.9343	0.9229	0.9469	0.7573	0.9480	0.9090	0.9283
Scaffold	0.9207	0.9297	0.9026	0.9474	0.6347	0.9528	0.8813	0.9170
FedSM	**0.9653**	**0.9489**	**0.9545**	**0.9551**	**0.9241**	**0.9560**	**0.9507**	**0.9527**

Table 9. Test Dice coefficient comparison of retinal disc segmentation.

Method	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Client Avg Dice	Global Dice
Centralized	0.8649	0.8033	0.8027	0.8507	0.7778	0.8793	0.8298	0.8507
Client 1 Local	0.8216	0.2306	0.5733	0.3793	0.2351	0.5621	0.4670	0.4675
Client 2 Local	0.1756	0.7810	0.0673	0.7184	0.1143	0.3637	0.3701	0.4511
Client 3 Local	0.7256	0.2807	0.8064	0.5621	0.2939	0.7333	0.5670	0.6068
Client 4 Local	0.3385	0.7749	0.2109	0.8491	0.1632	0.5842	0.4868	0.6024
Client 5 Local	0.4380	0.1000	0.3305	0.1560	0.7414	0.5380	0.3840	0.3952
Client 6 Local	0.7011	0.5360	0.6296	0.6886	0.3073	0.8752	0.6230	0.7198
FedAvg	0.8140	0.7949	0.7963	0.8495	0.7101	0.8795	0.8074	0.8402
FedProx	0.7822	0.7702	0.7864	0.8437	0.6132	0.8712	0.7778	0.8216
Scaffold	0.7554	0.7729	0.7405	0.8396	0.4995	0.8732	0.7469	0.8081
FedSM	**0.8610**	**0.8049**	**0.8186**	**0.8530**	**0.7724**	**0.8830**	**0.8322**	**0.8529**

Table 10. Test Dice coefficient comparison of retinal cup segmentation.

Unseen Client k	GM	PM1	PM2	PM3	PM4	PM5	PM6	Best γ, Dice
Client $k=6$	1.00	0	0	0	0	0	N/A	1, 0.8906
Client $k=5$	0.69	0.18	0	0	0.10	N/A	0.03	0.9, 0.4304
Client $k=4$	0.03	0	0.97	0	N/A	0	0	<0.95, 0.8870
Client $k=3$	0	0	0.57	N/A	0	0	0.43	<0.9, 0.8446
Client $k=2$	0	0	N/A	0	0.92	0.08	0	<1, 0.8409
Client $k=1$	0	N/A	0	1.00	0	0	0	<0.99, 0.8839

Table 11. (retinal segmentation, Dice = average of disc and cup Dice coefficients) Model selection frequency from the model selector when FL train with clients $\{1, 2, \cdots, 6\}/\{k\}$ and test on the **unseen** client $k \in \{1, 2, \cdots, 6\}$. From left to right, GM denotes the global model and PM denotes the personalized model $\{1, 2, \cdots, 6\}/\{k\}$. Here we choose the best γ.

Method/Unseen	Client 1	Client 2	Client 3	Client 4	Client 5	Client 6	Avg
Centralized	0.8842	0.8454	0.8214	0.8866	0.4064	0.8811	0.7875
FedAvg	0.8598	0.8313	0.8224	0.8551	0.4064	**0.8887**	0.7773
FedProx	0.8380	0.7856	0.8267	0.8746	0.4171	0.8784	0.7701
Scaffold	0.8085	0.7998	0.8211	0.8568	0.4121	0.8708	0.7615
FedSM	**0.8818**	0.8619	**0.8498**	**0.8901**	0.4118	0.8646	0.7933
FedSM-extra	0.8747	**0.8685**	0.8467	0.8794	**0.4265**	0.8809	**0.7963**

Table 12. (retinal segmentation, Dice = average of disc and cup Dice coefficients) Dice performance when FL train with clients $\{1, 2, \cdots, 6\}/\{k\}$ and test on the **unseen** client $k \in \{1, 2, \cdots, 6\}$.