FEDOPS: ORDER-SENSITIVE PARAMETER SANITIZATION FOR CROSS-CLIENT FEDERATED LEARNING IN MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, cross-client federated learning has received growing attention in medical image segmentation, owing to its superior performance and improved security guarantees compared with conventional server-based federated learning paradigms. However, existing methods suffer from two key challenges: 1) Severe data heterogeneity (non-IID) across clients often causes conflicting gradient directions during parameter exchanges, resulting in training oscillations that degrade convergence rate and final segmentation accuracy. 2) Critical parameter contamination during parameter transmission may cause interference parameters to overwhelm critical ones, resulting in incorrect pixel predictions. To overcome these issues, we propose a novel Order-sensitive Parameter Sanitization (FedOPS) framework for decentralized federated foundation models, which ensures more secure and stable training. Specifically, to promote stable convergence and effective knowledge propagation between different clients, we develop a consistency-aware client scheduling (CCS) strategy that orders clients based on feature representation similarity. To prevent critical parameter contamination, we introduce a cross-client parameter sanitization (CPS) mechanism that leverages historical update directions to selectively discard interfering updates. Extensive experiments on multiple benchmarks demonstrate that FedOPS consistently outperforms existing methods, highlighting its effectiveness for federated medical image segmentation.

1 INTRODUCTION

Federated Learning (FL) is a distributed machine learning paradigm that allows multiple decentralized clients to collaboratively train a shared model without exchanging raw data, thereby preserving data privacy Regulation (2018). In recent years, Federated learning has emerged as a promising approach in medical image segmentation, effectively addressing the reliance of most existing data-driven methods on large-scale centralized datasets Cai et al. (2018; 2021); Zhang et al. (2024b). At the same time, advances in large-scale foundation models Cheng et al. (2023) have attracted considerable attention for their potential in diverse clinical applications Tang et al. (2024), particularly medical image segmentation. Against this background, the integration of federated learning with large-scale foundation models has emerged as a promising paradigm for medical image segmentation, enabling privacy-preserving collaboration while effectively leveraging the strong representation capacity of foundation models across heterogeneous medical datasets. Existing federated medical image segmentation methods can be broadly categorized into two groups: server-based aggregation methods and cross-client federated learning methods.

In server-based aggregation approaches Wang et al. (2023); Liu et al. (2021); Jiang et al. (2023), clients first download a shared global model from a central server and perform local optimization using their private data. The resulting local updates or model parameters are then uploaded to the server, which aggregates them to produce an updated global model for the next communication round. For instance, Liu *et al.* Liu et al. (2024) integrate the SAM Ma et al. (2024); Kirillov et al. (2023) model with federated learning, demonstrating the potential of foundation models in federated medical image segmentation. While these methods have shown encouraging performance, server-based aggregation remains vulnerable to single-point failures, where server attacks or crashes can halt training and undermine robustness. In contrast, cross-client federated learning Xu et al. (2024);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

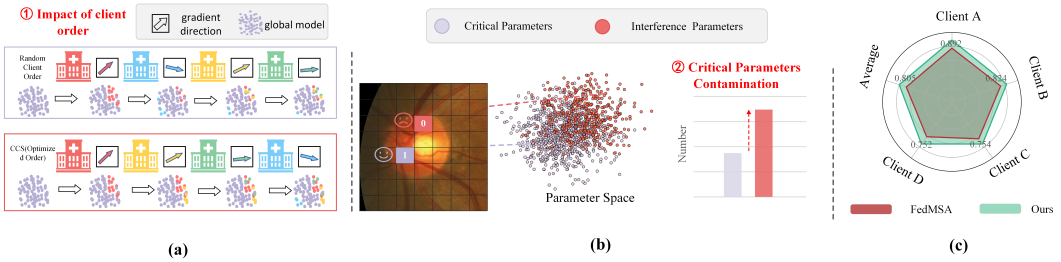

Figure 1: (a) Impact of client scheduling on model convergence. (b) The number of critical parameters for correct pixel prediction is much smaller than the interference ones, i.e., "critical parameter contamination" problem. (c) The performance comparison across different clients on the fundus dataset.

Kalra et al. (2023) methods enable direct communication among clients, avoiding the interruption issues inherent in server-based aggregation, and have therefore gained significant attention. For instance, Zhang Zhang et al. (2024a) overcome the catastrophic forgetting problem by leveraging cross-incremental collaborative distillation.However, all existing cross-client methods still face two major challenges during training. **The first challenge** is that the order in which clients participate in cross-client training can affect the consistency of knowledge propagation and the stability of convergence. Improper client scheduling may lead to underutilization or interference of client-specific knowledge, reducing overall performance (see Fig. 1(a)). **The second challenge** arises from critical parameters contamination during parameter transmission: parameters that correctly predict pixel categories contribute significantly to performance improvement, while interference parameters with incorrect predictions may even hinder it. Moreover, the number of critical parameters is typically much smaller than that of interference parameters, exacerbating the problem in practice (see Fig. 1(b)).

To address these challenges, we propose a novel Order-sensitive Parameter Sanitization framework for cross-client federated foundation models, dubbed FedOPS, which comprises two key components: consistency-aware client scheduling (CCS) and cross-client parameter sanitization (CPS). Specifically, unlike traditional server-based approaches, FedOPS performs training directly across clients, making the update order particularly important. To tackle this issue, consistency-aware client scheduling (CCS) arranges clients based on the similarity of their feature representations, ensuring that clients with similar data distributions are trained consecutively. This ordering promotes smoother and more coherent parameter updates while mitigating interference from heterogeneous data. Furthermore, FedOPS directly transmits trainable parameters across clients, which may cause critical parameter contamination, where interference updates overwhelm the relatively few critical parameters. To address this, cross-client parameter sanitization (CPS) evaluates the update consistency of each parameter over multiple training rounds. Parameters that consistently follow a stable direction are preserved, while those showing conflicting or irregular updates are masked out before transmission. By maintaining a parameter-level consistency score and applying a threshold-based mask, CPS selectively propagates only the critical and stable parameters to the next client, explicitly reducing interference and ensuring coherent knowledge transfer across clients. Extensive experiments on multiple benchmarks demonstrate that FedOPS significantly outperforms existing federated medical image segmentation methods. Our contributions can be summarized as follows:

- We propose a novel Order-sensitive Parameter Sanitization framework for decentralized federated foundation models that enables direct cross-client parameter transmission using the pretrained SAM model, preventing training interruptions without relying on a central server.
- We propose the FedOPS framework, which integrates **consistency-aware client scheduling** (CCS) to optimize the sequence of client updates and **cross-client parameter sanitization** (CPS) to preserve critical parameters and filter interference.
- Extensive experiments on three challenging benchmarks demonstrate that FedOPS significantly outperforms recent state-of-the-art methods across all clients in federated medical image segmentation.

2 Related Work

2.1 Server-based Federated Learning.

Federated learning has been widely applied to medical image segmentation to address privacy constraints. Early work such as FedAvg McMahan et al. (2017) lays the foundation for server-based model aggregation. Building on these advances, TM-FAL Yan et al. (2025) introduces a temporal model-based federated active learning framework, which leverages temporal information to guide client-side sample selection and model updating, aiming to improve learning efficiency under limited annotation budgets. FedDP Wang et al. (2023) and FedDG Liu et al. (2021) incorporate differential privacy and domain generalization mechanisms to enhance data privacy protection and model robustness under heterogeneous client distributions. FedFMS Liu et al. (2024) integrates foundation models like SAM Kirillov et al. (2023) into federated segmentation, demonstrating enhanced generalization across modalities. Jiang *et al.* Jiang et al. (2023) propose task-aware optimization to balance heterogeneous client updates, while ProxyFL Kalra et al. (2023) uses auxiliary proxy data to improve model consistency. These server-based approaches emphasize centralized aggregation, yet remain vulnerable to interruptions caused by server failures or attacks, motivating the cross-client training strategies.

2.2 Cross-client Federated Learning.

Cross-client , also known as fully decentralized federated learning removes the dependence on a central server by enabling direct peer-to-peer communication among clients. This paradigm also helps mitigate catastrophic forgetting, where local knowledge may be overwritten after global or cross-client updates. For instance, Zhang *et al.* Zhang et al. (2024a) leverage cross-incremental collaborative distillation to preserve locally learned information while enabling cross-client knowledge sharing. BlockDFL Qin et al. (2024), a blockchain-based fully decentralized FL framework, coordinates client updates via a PBFT-based voting and scoring mechanism to defend against poisoning attacks and enable scalable peer-to-peer collaboration without a trusted server. In graph learning scenarios, FedC4 Chen et al. (2025) introduces client-to-client collaboration with graph condensation to reduce communication cost and privacy risk by distilling private graph data into compact embeddings tailored for target clients. Other decentralized schemes explore gossip-based or peer-exchange communicationHegedűs et al. (2019); Onaizah et al. (2025) to further enhance robustness under dynamic client connectivity. While these approaches avoid server-related interruptions and help retain local knowledge, challenges such as critical parameters contamination and unstable convergence remain, which motivate our design of FedOPS with cross-client parameter sanitization and consistency-aware client scheduling.

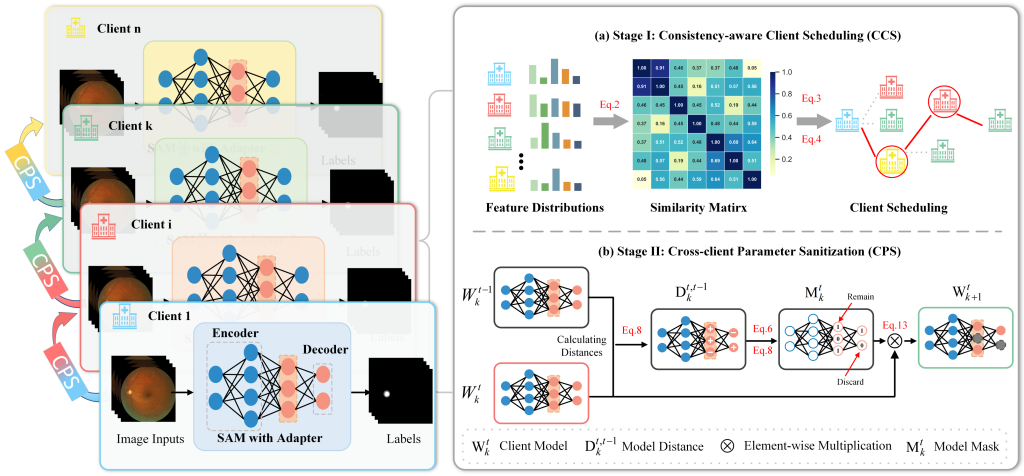

Figure 2: Architecture of the proposed FedOPS framework. FedOPS leverages a federated cross foundation model and combines the Critical Parameter Selection (CPS) and Coordinated Client Selection (CCS) to improve cross-client knowledge transfer.

3 METHODS

In this section, we provide a description of the implementation of the proposed FedOPS, as depicted in Figure 2.

3.1 PROBLEM FORMULATION AND MODEL SETUP

We consider a federated cross-client learning framework consisting of K clients and T communication rounds. Each client k possesses a private dataset $S_k = \{(x_n, y_n)\}_{n=1}^{N_k}$, where N_k denotes the number of local training samples. The objective is to collaboratively optimize the client models for medical image segmentation without exchanging raw data, which is formulated as

$$\min_{W_1,\ldots,W_K} \sum_{k=1}^{K} \mathcal{L}_k(W_k), \tag{1}$$

where $\mathcal{L}_k(\cdot)$ denotes the local segmentation loss of client k.

Unlike conventional server-based federated learning, FedOPS follows a decentralized cross-client training paradigm. In each communication round, clients are arranged according to the proposed Consistency-aware Client Scheduling (CCS) strategy. Each client initializes its model using the parameters received from the preceding client, performs local optimization on its private dataset, refines the updated parameters via the proposed Cross-client Parameter Sanitization (CPS), and forwards the sanitized parameters to the subsequent client. After all clients complete one training cycle, the next communication round is initiated.

For model instantiation, we adopt the Segment Anything Model (SAM) Kirillov et al. (2023) with parameter-efficient fine-tuning. Specifically, a lightweight segmentation head is appended after the decoder for multi-class prediction, and two adapter modules are inserted into each ViT Dosovitskiy (2020) encoder block. Accordingly, the model parameters of client k are denoted as $W_k = \{W_k^{enc}, W_k^a, W_k^{dec}\}$, where W_k^{enc}, W_k^a, and W_k^{dec} represent the encoder, adapters, and decoder, respectively. During federated optimization, the pretrained encoder remains frozen, while only the adapters and decoder are updated and transmitted across clients.

3.2 CONSISTENCY-AWARE CLIENT SCHEDULING

In cross-client federated learning, the order in which clients perform local training plays a crucial role in knowledge propagation and convergence stability, particularly when client data are heterogeneous. Random or arbitrary client sequences may introduce conflicting updates, slowing convergence or causing oscillations. To address this, we propose consistency-aware client scheduling, which arranges clients based on the similarity of their feature representations, ensuring smoother and more coherent parameter transitions across the network.

Formally, let $C = c_1, c_2, \ldots, c_N$ denote the set of clients, and $f_i \in \mathbb{R}^d$ represent the feature vector summarizing the local data characteristics of client c_i. We compute the pairwise cosine distance between clients:

$$D_{ij} = 1 - \frac{f_i \cdot f_j}{|f_i|_2 |f_j|_2}, \quad \forall i, j \in 1, \ldots, N, \tag{2}$$

where smaller D_{ij} values indicate higher similarity. The starting client c_s is selected as the one with minimal average distance to all other clients, representing the most central or representative client:

$$s = \arg\min_i \frac{1}{N} \sum_{j=1}^{N} D_{ij}. \tag{3}$$

Subsequent clients are scheduled iteratively in a greedy fashion. Let $\mathcal{V}t$ denote the set of clients already scheduled, and c_t the last scheduled client. The next client $ct + 1$ is chosen to minimize the distance to c_t among the remaining clients:

$$c_{t+1} = \arg \min_{c_k \in C \setminus \mathcal{V}_t} D(c_t, c_k), \tag{4}$$

the resulting ordered client sequence is then

$$\text{Order} = [c_s, c_{s_1}, c_{s_2}, \ldots, c_{s_{N-1}}]. \tag{5}$$

By arranging clients with similar characteristics consecutively, CCS promotes consistent parameter updates and mitigates interference caused by heterogeneous data distributions. When combined with CPS, CCS ensures that critical parameters propagate through the client network in a coherent and orderly manner, further improving convergence stability and overall performance. Integrating CCS into the federated cross foundation model allows each client to fully leverage the pretrained general knowledge of SAM while receiving parameter updates in an optimized sequence, thereby enabling a robust and efficient cross-client training process.

3.3 Cross-client Parameter Sanitization

Critical parameters tend to exhibit more stable update directions across training rounds, whereas interference parameters often undergo more diverse and inconsistent directional changes. Such directional consistency provides an effective criterion for characterizing the reliability of parameter updates based on their historical behavior. In particular, parameters that consistently move in the same direction across consecutive rounds are more likely to preserve coherent optimization information, while frequently changing directions may indicate unstable updates that can introduce conflicting information when transmitted across heterogeneous clients. Therefore, instead of blindly transferring all parameters, we selectively preserve parameters with strong historical directional consistency and suppress those with weak consistency. This strategy enables the transmitted model to retain informative parameter updates while reducing the propagation of potentially detrimental interference. Based on this observation, we propose a cross-client parameter sanitization strategy. Specifically, without relying on the internal structure of the model, we represent the parameters of client k at round t as:

$$W_k^t = [w_{k,1}^t, w_{k,2}^t, w_{k,3}^t, \ldots, w_{k,L}^t], \tag{6}$$

where L denotes the total number of parameters. We then calculate the model distance $D_k^{t,t-1}$ from the previous round of the current client, which can be formulated as:

$$D_k^{t,t-1} = [d_{k,1}^{t,t-1}, d_{k,2}^{t,t-1}, d_{k,3}^{t,t-1}, \ldots, d_{k,L}^{t,t-1}], \tag{7}$$

$$D_k^{t,t-1} = [(w_{k,1}^t - w_{k,1}^{t-1}), \ldots, (w_{k,L}^t - w_{k,L}^{t-1})]. \tag{8}$$

After obtaining $D_k^{t,t-1}$, we can determine whether the update direction of all model parameters, compared to the previous round, is in the positive direction (greater than 0) or in the negative direction (less than 0). Additionally, considering that updates from previous rounds should also be taken into account for parameter pruning, we define the consistency of parameters C_k^t as:

$$C_k^t = [c_{k,1}^t, c_{k,2}^t, c_{k,3}^t, \ldots, c_{k,L}^t], \tag{9}$$

$$C_k^t \leftarrow \frac{c_{k,l}^{t-1} * (t-1) + I(d_{k,l}^{t,t-1} \geq 0)}{t}, \tag{10}$$

where I$(\cdot)$ is the indicator function, $c_{k,l}^t$ represents the consistency value of the l-th parameter of the k-th client in the t-th round. Finally, we adjust the value of each item in C_k^t to better represent the distance of the parameters moving in the positive or negative direction. Additionally, we introduce a threshold θ to generate a parameter mask M_k^t ,which is used to filter out parameters with weak consistency, while retaining those with strong consistency:

$$C_k^t = \begin{cases} c_{k,l}^t & \text{if} \quad d_{k,l}^{t,t-1} \geq 0, \\ 1 - c_{k,l}^t & \text{otherwise} \end{cases}, \tag{11}$$

$$M_k^t = [m_{k,1}^t, m_{k,2}^t, \ldots, m_{k,L}^t] = \begin{cases} 1 & \text{if} \quad c_{k,l}^t \geq \theta, \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

The parameters W_{k+1}^t passed to the next client are derived from the following equation:

$$W_{k+1}^t = W_k^t \otimes M_k^t, \tag{13}$$

where $\otimes$ denotes element-wise multiplication.

3.4 Loss Function

Follow the previous work Liu et al. (2024), the loss function we use is the cross-entropy loss, which can be defined as follows:

$$L_{CE} = -\sum_{i=1}^{S}\sum_{j=1}^{Z} \frac{1}{N} \cdot y_{ij} \log\left(p_{ij}\right), \tag{14}$$

where S is the number of pixels, Z is the number of categories. p_{ij} is the prediction probability that pixel i belongs to category j, y_{ij} represents whether pixel i belongs to category j. If pixel i belongs to category j, then y_{ij} is 1, otherwise 0.

4 Experiments

4.1 Experiment Setup

Datasets & Metrics: Following previous work Liu et al. (2024), we evaluate our method on three publicly available federated medical image segmentation datasets, including prostate cancer Litjens et al. (2014); Lemaître et al. (2015), fundus Sivaswamy et al. (2015); Orlando et al. (2020); Bajwa et al. (2020); Zhang et al. (2010), and nuclei segmentation datasets Gamper et al. (2020); Verma et al. (2021); Naylor et al. (2018); Kumar et al. (2017). These datasets involve diverse imaging modalities and heterogeneous non-IID client distributions, and are partitioned into multiple clients to simulate realistic federated learning scenarios with institutional heterogeneity. All images are resized to 1024×1024×3, and the label crop size is set to 256×256. We adopt two widely used segmentation metrics, Dice similarity coefficient (Dice) and Intersection over Union (IOU) Xu et al. (2024), to quantitatively evaluate segmentation performance.

Implementation Details: In our experiments, each client corresponds to a dataset from a different hospital. We adopt a leave-one-client-out strategy: in each round, one client serves as an unseen test set, while data from the remaining clients are split 9:1 into training and validation for federated cross-training. This ensures evaluation on entirely unseen institutional data. The final test weights are taken from each client after the last global epoch. All clients are initialized with SAM's pre-trained ViT-B, optimized with Adam (learning rate 0.0001, batch size 2). We set the threshold $\theta = 0.1$ and train for 100 global rounds, with one local epoch per client per round.

Model	Client A		Client B		Client C		Client D		Client E		Client F		Average	
	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU
FedUNet	0.498	0.498	0.684	0.645	0.034	0.023	0.649	0.590	0.671	0.640	0.563	0.542	0.516	0.490
FedCross	0.622	0.579	0.653	0.602	0.615	0.615	0.272	0.271	0.547	0.503	0.450	0.449	0.527	0.503
FedSAM	0.810	0.774	0.841	0.808	**0.798**	**0.768**	0.837	0.792	0.785	0.754	0.844	0.809	0.819	0.784
FedMSA	0.769	0.737	0.809	0.777	0.754	0.720	0.821	0.771	0.795	0.765	0.859	0.820	0.801	0.765
TM-FAL	0.812	0.779	0.845	0.822	0.782	0.753	0.831	0.785	0.789	0.761	0.862	0.828	0.820	0.788
Ours	**0.818**	**0.783**	**0.856**	**0.823**	0.778	0.743	**0.846**	**0.795**	**0.833**	**0.804**	**0.876**	**0.840**	**0.835**	**0.798**

Table 1: Quantitative comparisons on the Prostate Cancer Dataset. Bold values indicate the best performance.

Model	Client A		Client B		Client C		Client D		Average	
	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU
FedUNet	0.848	0.743	0.794	0.667	0.550	0.427	0.677	0.529	0.717	0.592
	0.840	0.733	0.830	0.722	0.449	0.346	0.779	0.650	0.725	0.613
FedCross	0.841	0.732	0.792	0.667	0.588	0.462	0.639	0.492	0.715	0.588
	0.826	0.712	0.791	0.671	0.495	0.392	0.754	0.615	0.717	0.584
FedSAM	0.869	0.772	0.816	0.698	0.717	0.602	0.715	0.571	0.779	0.661
	0.873	0.781	0.846	0.748	0.556	0.456	0.777	0.643	0.763	0.657
FedMSA	0.860	0.760	0.822	0.705	0.733	0.614	0.692	0.545	0.777	0.656
	0.858	0.763	0.850	0.753	0.560	0.463	0.769	0.632	0.759	0.653
TM-FAL	0.876	0.785	0.817	0.701	0.742	0.629	0.722	0.581	0.789	0.674
	0.865	0.772	0.848	0.743	0.574	0.478	0.790	0.655	0.769	0.662
Ours	**0.892**	**0.809**	**0.824**	**0.709**	**0.754**	**0.644**	**0.752**	**0.617**	**0.805**	**0.695**
	0.872	**0.782**	**0.854**	**0.760**	**0.588**	**0.492**	**0.806**	**0.682**	**0.780**	**0.679**

Table 2: Quantitative comparisons on the Fundus Dataset. Each table cell contains two numbers: the top one is the OD result, and the bottom one is the OC result.

Model	Client A		Client B		Client C		Client D		Client E		Client F		Client G		Average	
	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU	Dice	IOU
FedUNet	0.742	0.618	0.781	0.655	0.717	0.598	0.777	0.645	0.559	0.404	0.703	0.548	0.678	0.519	0.708	0.569
FedCross	0.760	0.636	0.782	0.658	0.730	0.613	0.791	0.661	0.571	0.417	0.700	0.554	0.629	0.467	0.709	0.572
FedSAM	0.810	0.698	0.807	0.693	0.765	0.659	0.832	0.717	0.623	0.472	0.776	0.643	0.746	0.602	0.765	0.640
FedMSA	0.806	0.694	0.802	0.688	0.763	0.655	0.805	0.682	**0.640**	**0.490**	0.730	0.600	0.741	0.598	0.755	0.630
TM-FAL	0.809	0.699	0.815	0.704	0.780	0.675	0.820	0.700	0.634	0.483	0.752	0.620	0.756	0.618	0.767	0.643
Ours	**0.820**	**0.713**	**0.831**	**0.725**	**0.796**	**0.695**	**0.832**	**0.717**	0.628	0.477	**0.780**	**0.651**	**0.772**	**0.637**	**0.780**	**0.659**

Table 3: Quantitative comparisons on the Nuclei Dataset.

4.2 Performance Comparison

In this paper, we compare our method with five approaches: FedUNet, FedCross, FedMSA, FedSAM, and TM-FAL Yan et al. (2025). FedUNet, TM-FAL, and FedCross use UNet Ronneberger et al. (2015) or SAM backbones with FedAvg aggregation, while FedMSA fine-tunes only adapters Wu et al. (2023) and the decoder and FedSAM fine-tunes all SAM parameters. All methods are server-based except FedCross, which uses a cross-training paradigm.

Quantitative Comparison: Table 1 shows that FedOPS achieves state-of-the-art performance across most clients in the Prostate Cancer Dataset. For client C, FedOPS slightly underperforms FedSAM, likely because FedSAM updates all model parameters, whereas FedOPS adopts parameter-efficient fine-tuning with only a subset of parameters updated. Nevertheless, FedOPS surpasses other baselines, achieving average Dice and IOU scores of 0.835 and 0.798, respectively. This improvement is mainly attributed to CCS, which reduces feature-level heterogeneity by optimizing client update order, and CPS, which preserves critical parameters by filtering conflicting updates during transmission.

To further evaluate FedOPS, we report results on the Fundus and Nuclei datasets in Table 2 and Table 3. On the Fundus dataset, FedOPS achieves the best results across all four clients, with average Dice and IoU scores of 0.805 and 0.695. For optic cup (OC) segmentation, FedOPS improves Dice by 2.1% over the previous best baseline FedSAM (0.780 vs. 0.763), while maintaining competitive optic disc (OD) performance. The variation on client C may result from its smaller sample size and distinct image characteristics, yet FedOPS remains superior on other clients.

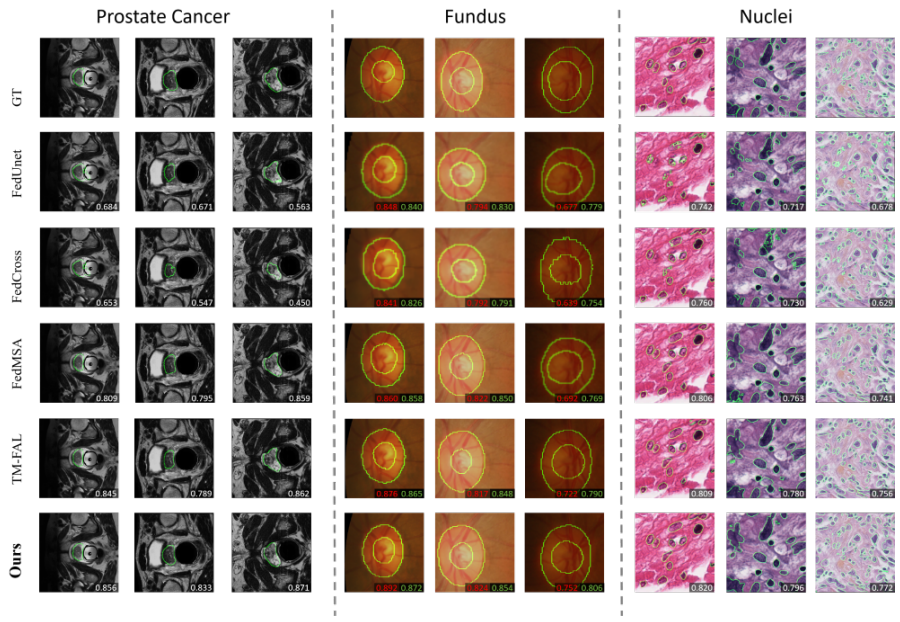

Figure 3: Qualitative comparisons on three datasets. The green curve in the image represents the contour of the segmentation mask, with the Dice score annotated in the bottom-right corner. The Dice score for the optic disc is shown in red, while the score for the optic cup is in green.

On the Nuclei dataset, FedOPS achieves average Dice and IoU scores of 0.780 and 0.659 across seven clients, outperforming TM-FAL by 1.3% and 1.6%, respectively. The robust performance across clients demonstrates FedOPS's resilience to heterogeneous data distributions.

Qualitative Comparison: Fig. 3 presents visual comparisons between our method and several recent state-of-the-art approaches across three datasets. Our approach effectively captures intricate anatomical structures, including small ducts, fine vessels, and subtle optic cup boundaries, which are frequently blurred or overlooked by baseline methods. By leveraging cross-client parameter transmission, FedOPS facilitates comprehensive knowledge sharing among clients, while its consistency-aware client scheduling selectively filters out unstable or noisy updates. This combination not only mitigates oscillatory behavior during training but also promotes smoother and more stable convergence across heterogeneous datasets. Importantly, the qualitative improvements illustrated in Fig. 3 are consistent with quantitative metrics such as Dice and IoU, demonstrating that the enhanced visual segmentation quality directly translates into measurable performance gains. Collectively, these results highlight that FedOPS achieves superior segmentation accuracy while maintaining robust and consistent performance across diverse clients, underscoring its practical applicability for real-world federated medical image analysis.

4.3 Ablation Study and Analysis

Ablation Study of Components: Table 4 evaluates the contributions of key FedOPS components on the Fundus dataset. Removing either CPS or CCS degrades performance, demonstrating that both components are essential for effective cross-client knowledge transfer. We further analyze the impact of client update orders in CCS (Fig. 4) and parameter selection strategies in CPS (Fig. 5). The results show that proper client ordering improves knowledge propagation and convergence stability, while CPS preserves critical parameter updates and enhances segmentation performance.

CPS	CCS	Dice	IOU
✗	✗	77.7	65.6
✓	✗	79.2	68.8
✗	✓	78.6	67.4
✓	✓	**80.5**	**69.5**

Table 4: Comparison of average OD Dice across all clients on the Fundus dataset for different components of FedOPS.

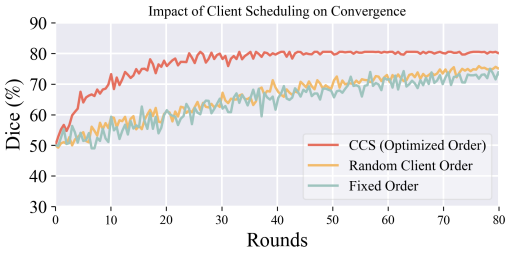

Figure 4: Impact of Client Scheduling on Convergence.

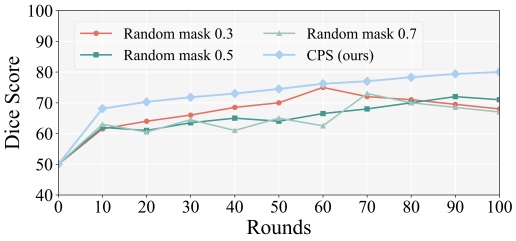

Figure 5: Comparison of random masking and CPS for critical parameter selection.

Backbone	A	B	C	D	Average
None	0.814	0.735	0.681	0.667	0.724
SAM-B	0.892	0.824	0.754	0.752	0.805
SAM-L	**0.907**	**0.831**	**0.783**	**0.779**	**0.825**

Table 5: Different backbone initial weights. "None" indicates no pretrained weights.

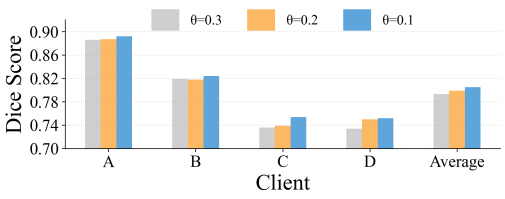

Figure 6: Analysis of hyper-parameter θ.

Ablation Study of the Different Initialization Weights: Table 5 shows the performance variations with different initialization weights, illustrating that using pretrained weights can effectively improve segmentation accuracy. However, larger pretrained models, while achieving better performance, also introduce increased computational overhead and consequently slow down the training process.

Parameter Analysis: Fig.6 analyzes the hyperparameter θ and its effect on the model's performance. A larger θ, although enforcing consistency over a greater number of parameters, may suppress the transmission of critical parameters, which in turn leads to performance degradation.

Efficiency and Model Complexity Analysis: Table 6 compares the segmentation performance, communication efficiency, and model complexity of different methods on the Prostate dataset. Our method achieves the highest Dice score of 0.835 while requiring only 70 communication rounds, fewer than all competing methods. This improvement is achieved without increasing the model size or introducing additional computational overhead during training. Meanwhile, it maintains a lightweight model with only 1.2M parameters, demonstrating an effective balance between segmentation accuracy, communication efficiency, and model complexity.

Method	**Prostate**		
	Dice	**Rounds**	**Params (M)**
FedUNet	0.516	88	1.2
FedCross	0.527	74	1.3
FedMSA	0.801	88	1.8
TM-FAL	0.820	76	1.5
Ours	**0.835**	**70**	**1.2**

Table 6: Performance on the Prostate Dataset.

5 Conclusion

This paper presents the Order-sensitive Parameter Sanitization framework (FedOPS), a federated foundation model for medical image segmentation integrating cross-client knowledge refinement and selective parameter updating. It applies foundation models in a server-free federated cross-learning framework, enabling efficient training across heterogeneous clients. Consistency-aware client scheduling optimizes the update order to enhance knowledge propagation, while cross-client parameter sanitization preserves critical parameters and mitigates contamination. Extensive experiments and ablation studies on three challenging medical imaging benchmarks demonstrate the robustness and generalization ability of the proposed method.

REFERENCES

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, and Sheraz Dengel. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *International Joint Conference on Neural Networks*, pp. 1–7. IEEE, 2020.

Qing Cai, Huiying Liu, Sanping Zhou, Jingfeng Sun, and Jing Li. An adaptive-scale active contour model for inhomogeneous image segmentation and bias field estimation. *Pattern Recognition*, 82: 79–93, 2018.

Qing Cai, Yiming Qian, Sanping Zhou, Jinxing Li, Yee-Hong Yang, Feng Wu, and David Zhang. Avlsm: Adaptive variational level set model for image segmentation in the presence of severe intensity inhomogeneity and high noise. *IEEE Transactions on Image Processing*, 31:43–57, 2021.

Zekai Chen, Xunkai Li, Yinlin Zhu, Rong-Hua Li, and Guoren Wang. Fedc4: Graph condensation meets client-client collaboration for efficient and private federated graph learning. *arXiv preprint arXiv:2504.14188*, 2025.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.

István Hegedűs, Gábor Danner, and Márk Jelasity. Gossip learning as a decentralized alternative to federated learning. In José Pereira and Laura Ricci (eds.), *Distributed Applications and Interoperable Systems*, pp. 74–90, Cham, 2019. Springer International Publishing.

Meirui Jiang, Hongzheng Yang, Chen Cheng, and Qi Dou. Iop-fl: Inside-outside personalization for federated medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(7):2106–2117, 2023.

Shivam Kalra, Junfeng Wen, Jesse C Cresswell, and Hamid R Volkovs. Decentralized federated learning through proxy model sharing. *Nature Communications*, 14(1):2899, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.

Guillaume Lemaître, Robert Martí, Joan C Freixenet, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in Biology and Medicine*, 60:8–31, 2015.

Geert Litjens, Robert Toth, Wendy Van De Ven, Sjoerd Hoeks, Bram Van Ginneken, Gwenael Vincent, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021.

Yuxi Liu, Guibo Luo, and Yuesheng Zhu. Fedfms: Exploring federated foundation models for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pp. 283–293. Springer, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38 (2):448–459, 2018.

Ameer N. Onaizah, Yuanqing Xia, and Khurram Hussain. Fl-sicnn: An improved brain tumor diagnosis using siamese convolutional neural network in a peer-to-peer federated learning approach. *Alexandria Engineering Journal*, 114:1–11, 2025. ISSN 1110-0168. doi: https://doi.org/10.1016/j.aej.2024.11.063. URL https://www.sciencedirect.com/science/article/pii/S1110016824015345.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020.

Zhen Qin, Xueqiang Yan, Mengchu Zhou, and Shuiguang Deng. Blockdfl: A blockchain-based fully decentralized peer-to-peer federated learning framework. In *Proceedings of the ACM Web Conference*, 2024.

Protection Regulation. General data protection regulation. *Intouch*, 25:1–5, 2018.

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*, pp. 234–241. Springer, 2015.

Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015.

Jiao Tang, Yagao Yue, Peng Wan, Mingliang Wang, Daoqiang Zhang, and Wei Shao. Osal-nd: Open-set active learning for nucleus detection. In *Medical Image Computing and Computer Assisted Intervention*, pp. 351–361. Springer, 2024.

Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, et al. Monusac2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Transactions on Medical Imaging*, 40(12):3413–3423, 2021.

Jiacheng Wang, Danail Jin, and Liansheng Wang. Feddp: Dual personalization in federated medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):297–308, 2023.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

Xuanang Xu, Hannah H Deng, Tianyi Chen, Tianshu Kuang, Joshua C Barber, Daeseung Kim, Jaime Gateno, James J Xia, and Pingkun Yan. Federated cross learning for medical image segmentation. In *Medical Imaging with Deep Learning*, pp. 1441–1452. PMLR, 2024.

Yunlu Yan, Chun-Mei Feng, Yuexiang Li, Jinheng Xie, Jun Chen, Mohamed Elhoseiny, Ming Hu, Kaishun Wu, and Lei Zhu. Temporal Model-Based Federated Active Medical Image Classification . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15973. Springer Nature Switzerland, September 2025.

Fan Zhang, Huiying Liu, Qing Cai, Chun-Mei Feng, Binglu Wang, Shanshan Wang, Junyu Dong, and David Zhang. Federated cross-incremental self-supervised learning for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024a.

Fan Zhang, Huiying Liu, Xiaojun Duan, Binglu Wang, Qing Cai, Huafeng Li, Junyu Dong, and David Zhang. Dslsm: Dual-kernel-induced statistic level set model for image segmentation. *Expert Systems with Applications*, 242:122772, 2024b.

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Jun Lee, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3065–3068. IEEE, 2010.

A APPENDIX

To provide a more comprehensive explanation of our method, this supplementary material includes detailed information on various aspects of our approach:

- Pseudocode algorithm.
- Advantage discussions.
- External validation.
- More comparison results.

Detailed Algorithm

In Algorithm7, we summarize the detailed cross-client training procedure of our proposed FedOPS. Specifically, W_k^t denotes the trainable parameters of client k at communication round t, and C_k^t represents the consistency values of each parameter used in the Cross-client Parameter Sanitization (CPS) module. We discuss the advantages and main differences of our proposed method over Temporal Model-based Federated Active Learning (TM-FAL) Yan et al. (2025). TM-FAL estimates uncertainty by integrating feature differences and prediction confidence from temporal local and global models, and uses pseudo-label-based grouping to mitigate class imbalance. The main differences between TM-FAL and our proposed FedOPS are threefold: a) TM-FAL relies on a server-based paradigm, where clients depend on a central server for model aggregation, and training may be interrupted if the server fails. In contrast, FedOPS adopts a cross-client training paradigm, allowing direct parameter transmission among clients and improving robustness. b) TM-FAL captures local-global differences using temporal uncertainty. In contrast, FedOPS explicitly handles feature-level heterogeneity by using Consistency-aware Client Scheduling (CCS) to optimize the sequence of client updates, ensuring consistent knowledge propagation and stable convergence. c) TM-FAL does not address interference among parameters during cross-client transmission. In contrast, FedOPS directly transmits trainable parameters across clients, which may lead to critical parameter contamination. The relatively few critical parameters that contribute positively to prediction can be overwhelmed by interfering updates, reducing model performance. To mitigate this challenge, we propose Cross-client Parameter Sanitization (CPS), which leverages historical update directions to selectively discard interfering updates, thereby preserving the contribution of critical parameters, stabilizing optimization, and enhancing model accuracy.

Algorithm 1: FedOPS: Cross-Client Training with CCS and CPS

Require: a) Set of clients $C = \{c_1, \ldots, c_N\}$
b) Total communication rounds T
c) Consistency threshold θ for CPS
d) Initial client parameters W_k^0 for $k = 1, \ldots, N$
e) Feature vectors f_k summarizing local data

Input : None (all operations use local client data internally)

Output : Trained cross-client parameters W_k^T

for $t \leftarrow 1$ *to* T **do**

`// Consistency-aware Client Scheduling (CCS)`

Compute pairwise cosine distance D_{ij} between all clients;

Select starting client $c_s = \arg\min_i \frac{1}{N} \sum_j D_{ij}$;

Initialize scheduled set $\mathcal{V} \leftarrow \{c_s\}$;

for *each remaining client* $k \in C \setminus \mathcal{V}$ **do**

Select next client $c_{next} = \arg\min_{c_k \in C \setminus \mathcal{V}} D(c_{last}, c_k)$;

$\mathcal{V} \leftarrow \mathcal{V} \cup \{c_{next}\}$;

Ordered sequence: Order $= [c_s, \ldots, c_{last}]$;

`// Cross-client Parameter Sanitization (CPS)`

for *client* k *in Order* **do**

Compute parameter differences: $D_k^{t,t-1} = W_k^t - W_k^{t-1}$;

Update parameter consistency: $C_k^t \leftarrow \frac{C_k^{t-1} * ((t-1) + I(D_k^{t,t-1} \geq 0))}{t}$;

Adjust consistency values per parameter;

if $d_{k,l}^{t,t-1} \geq 0$ **then**

$c_{k,l}^t \leftarrow C_{k,l}^t$

else

$c_{k,l}^t \leftarrow 1 - C_{k,l}^t$

Generate parameter mask;

if $c_{k,l}^t \geq \theta$ **then**

$m_{k,l}^t \leftarrow 1$

else

$m_{k,l}^t \leftarrow 0$

Apply mask: $W_{k+1}^t \leftarrow W_k^t \otimes M_k^t$;

Return : W_k^T

External Validation

B DATASET INFORMATION

Prostate Cancer: We use data from six different sources as individual clients, constructing clients A through F for the prostate segmentation Litjens et al. (2014); Lemaître et al. (2015) (the number of 3D volumes are 12, 12, 30, 30, 13, and 19, respectively).

Fundus: The optic disc (OD) and optic cup (OC) segmentation task with retinal fundus images from 4 different institutions Sivaswamy et al. (2015); Orlando et al. (2020) (the number of samples are: 101, 1200, 650, and 1020, respectively).

Method	Prostate		Fundus OD		Fundus OC		Nuclei	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
TM-FAL	0.820	0.788	0.789	0.674	0.769	0.662	0.767	0.643
Ours	**0.835**	**0.798**	**0.805**	**0.695**	**0.780**	**0.679**	**0.780**	**0.659**

Table A: Dice and IoU results for TM-FAL and our method across Prostate, Fundus (OD/OC), and Nuclei datasets.

Nuclei: We use nuclei images from different institutions Gamper et al. (2020); Verma et al. (2021) for the nuclei segmentation task, dividing the data into a total of 7 clients (the number of samples are: 148, 128, 158, 163, 208, 50, and 32, respectively).

C External Validation Results

We perform external validation on three datasets, namely Prostate, Fundus, and Nuclei, which were not seen during the training phase, to evaluate the generalizability and robustness of our proposed FedOPS. In particular, for the Fundus dataset, we report metrics separately for the optic disc (OD) and optic cup (OC). Table A summarizes the comparison between TM-FAL Yan et al. (2025) and our method. It can be observed that TM-FAL suffers from performance degradation on these unseen datasets, especially on the Nuclei dataset. In contrast, our method consistently achieves higher Dice and IoU scores across all datasets. These improvements can be mainly attributed to the proposed cross-client training paradigm, including Consistency-aware Client Scheduling (CCS) and Cross-client Parameter Sanitization (CPS), which effectively handle feature heterogeneity and critical parameter contamination, thereby enhancing the generalization ability of the trained models.

D More Qualitative Results

To further demonstrate the effectiveness of our FedOPS, we provide qualitative comparisons with FedUnet, FedCross, FedMSA, TM-FAL, and ground-truth (GT) masks. For evaluation, each client's dataset is treated as an unseen test set, while the remaining clients' data are split 9:1 for federated cross-training, reflecting generalization to unseen clients. Figures A and B show two Nuclei images; Figures C, D, and E present three Prostate volumes; and Figures F and G display Fundus images with optic disc and optic cup (OC). Across datasets, FedOPS consistently produces more accurate and visually coherent segmentation masks than competing methods, especially in regions with weak boundaries or densely packed structures. These results highlight the effectiveness of our cross-client training paradigm, including Consistency-aware Client Scheduling (CCS) and Cross-client Parameter Sanitization (CPS), in improving generalization and robustness on unseen clients.

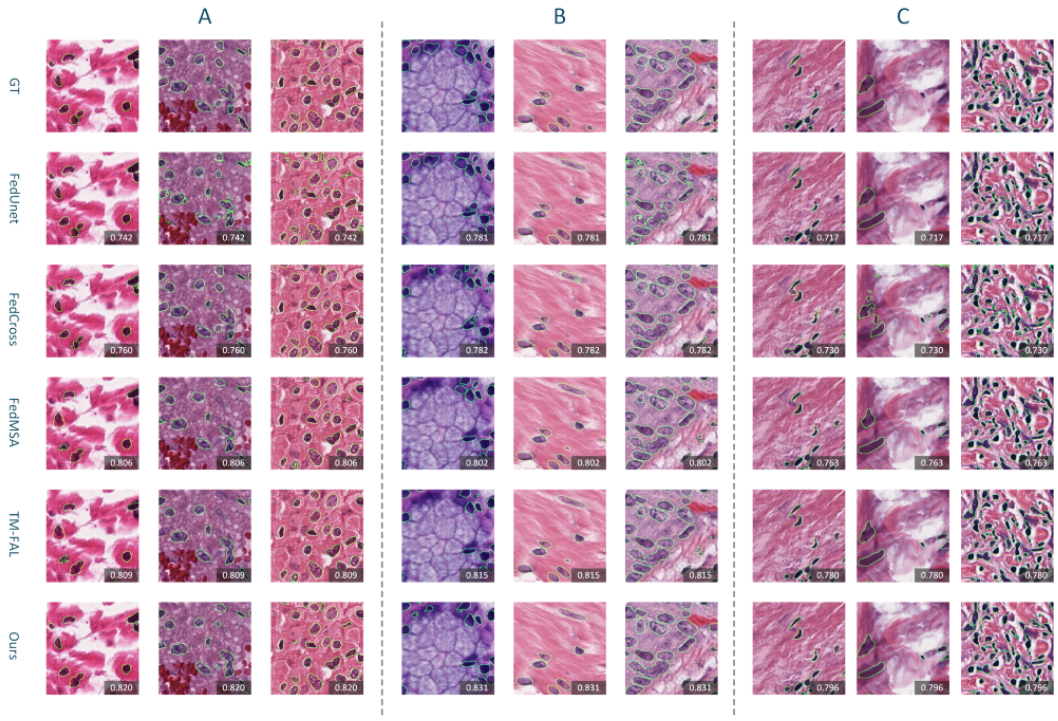

Figure A: Qualitative segmentation comparison on a Nuclei image from test clients A, B, and C. From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice value for each method is indicated in the bottom-right corner of each mask.

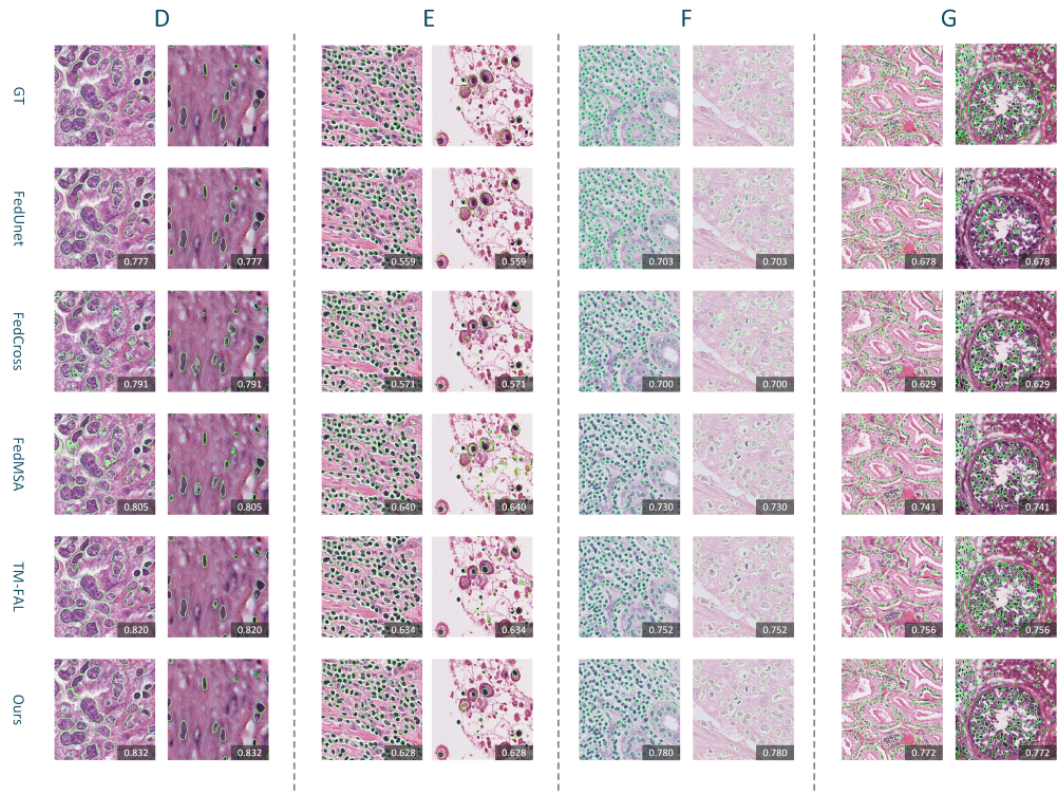

Figure B: Qualitative segmentation comparison on a Nuclei image from test clients D, E, F, and G. From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice value for each method is indicated in the bottom-right corner of each mask.

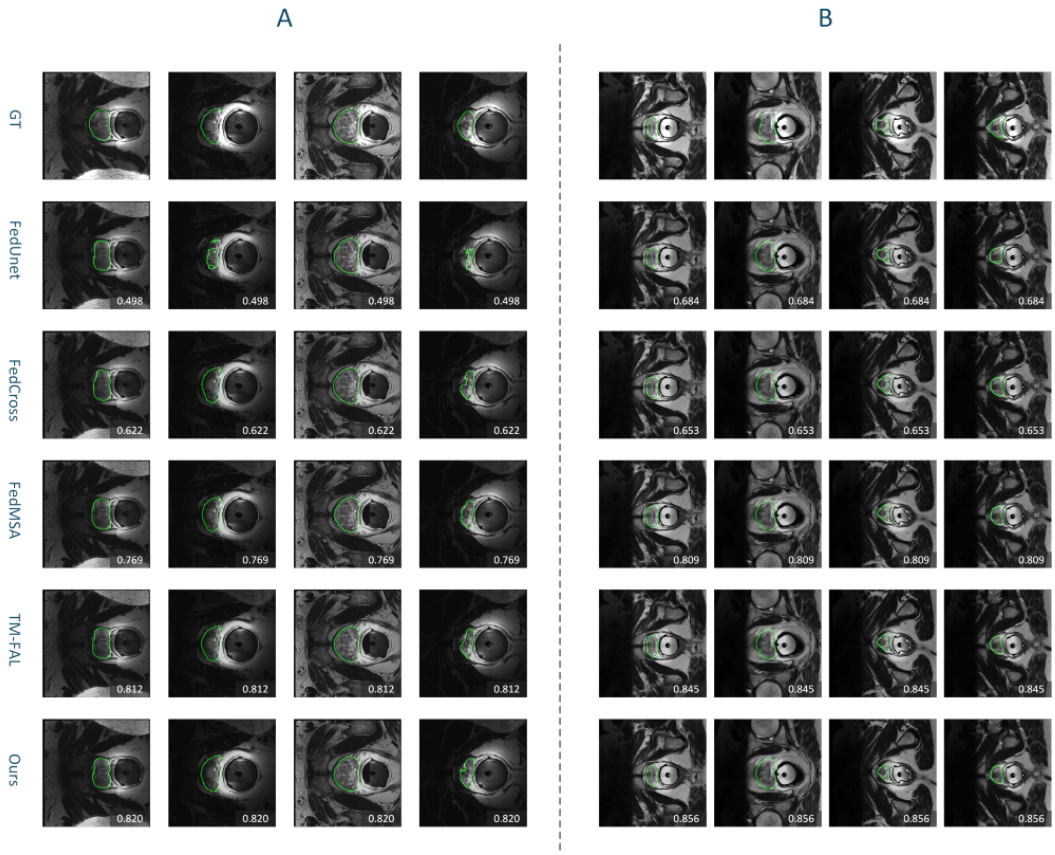

Figure C: Qualitative segmentation comparison on a Prostate volume from test clients A and B. From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice value for each method is indicated in the bottom-right corner of each mask.

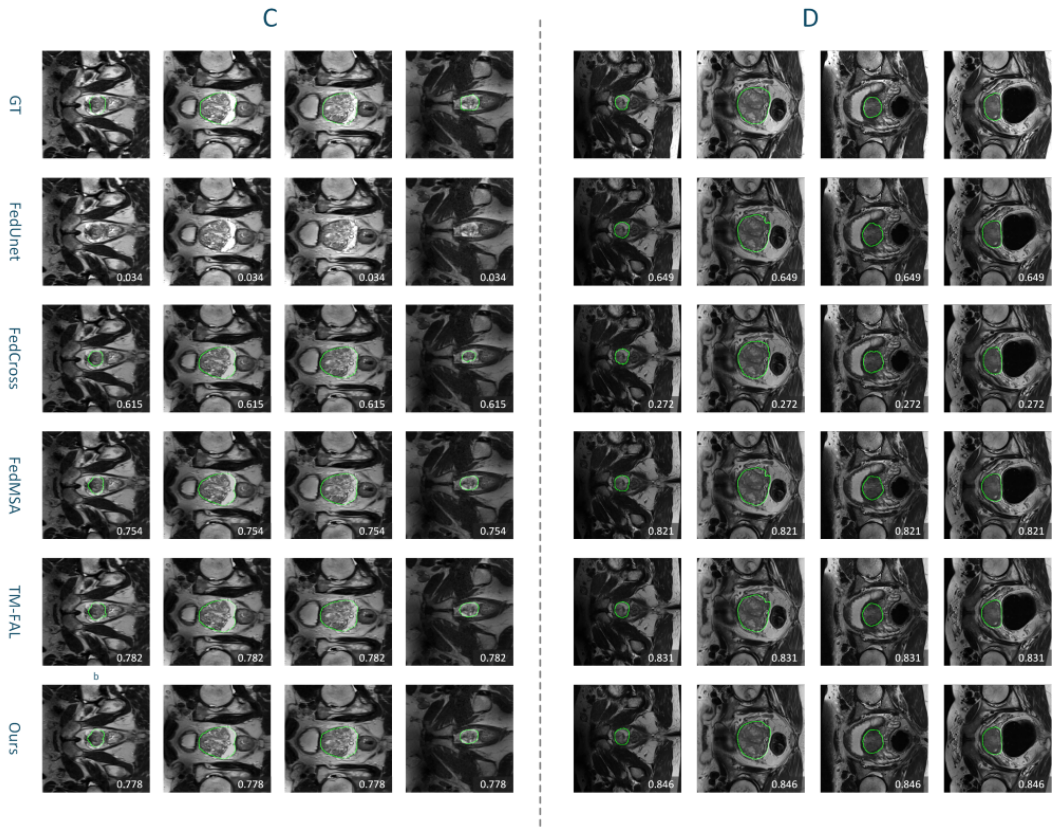

Figure D: Qualitative segmentation comparison on a Prostate volume from test clients C and D. From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice value for each method is indicated in the bottom-right corner of each mask.

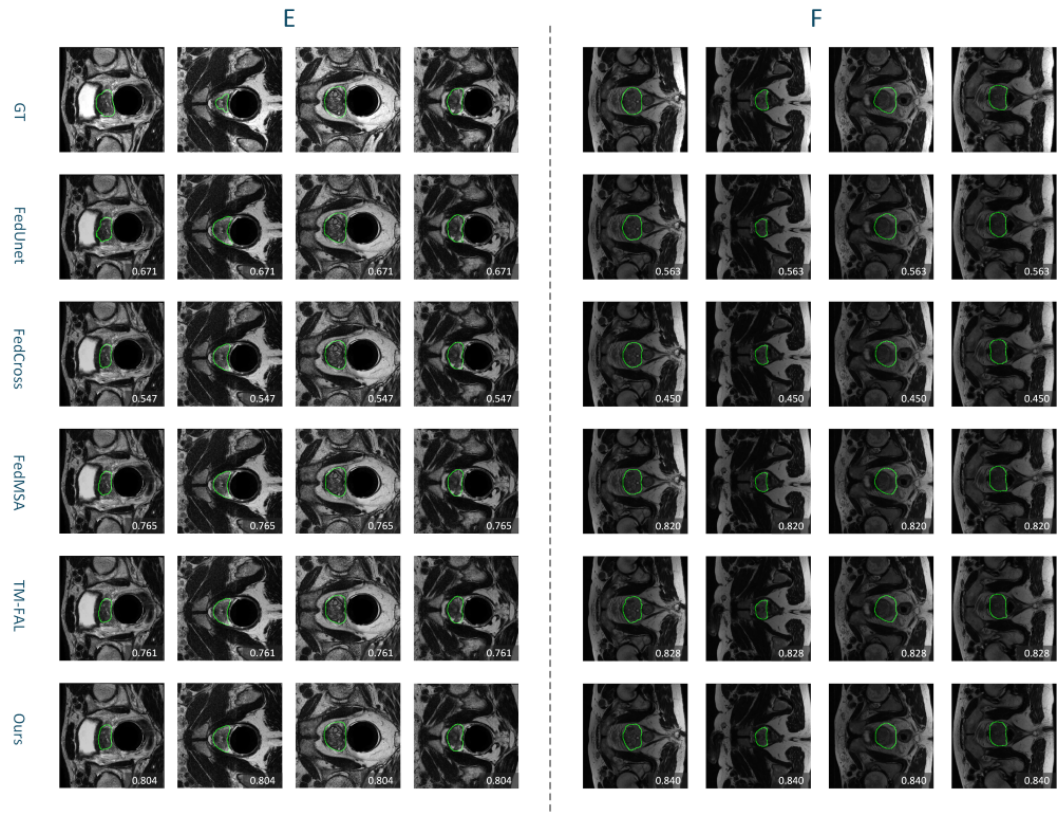

Figure E: Qualitative segmentation comparison on a Prostate volume from test clients E and F. From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice value for each method is indicated in the bottom-right corner of each mask.

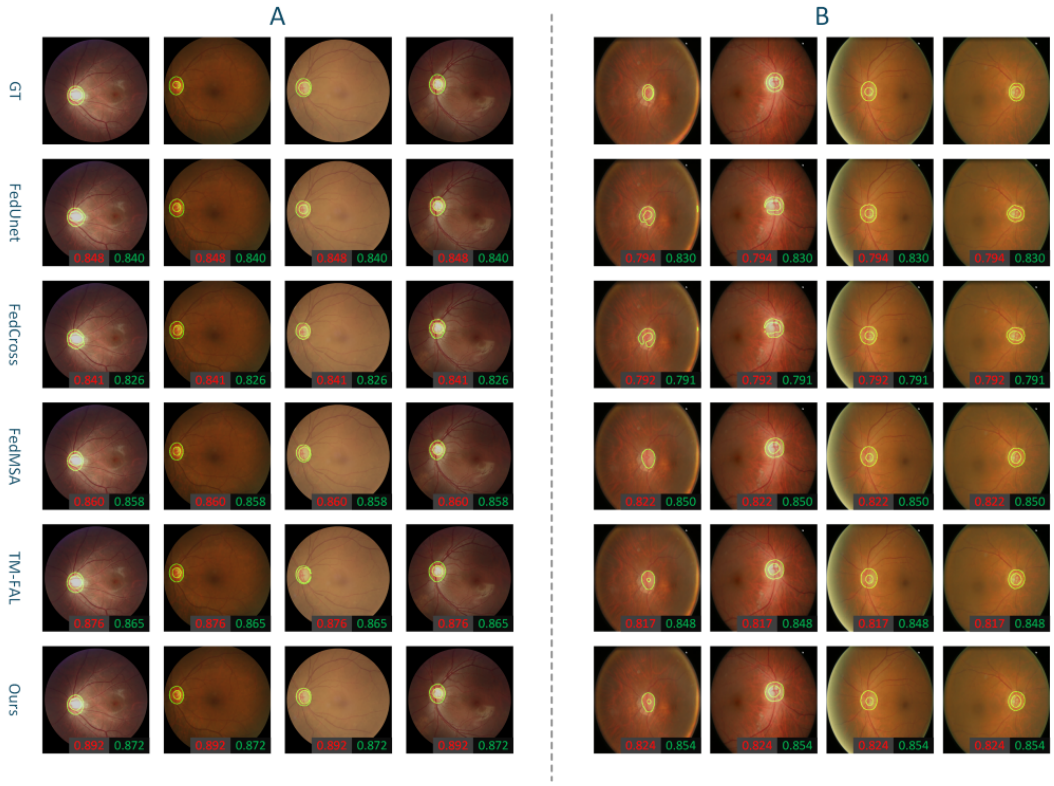

Figure F: Qualitative segmentation comparison on a Fundus image from test clients A and B, highlighting optic disc (OD) and optic cup (OC). From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice values for OD (red) and OC (green) are indicated in the bottom-right corner of each mask.

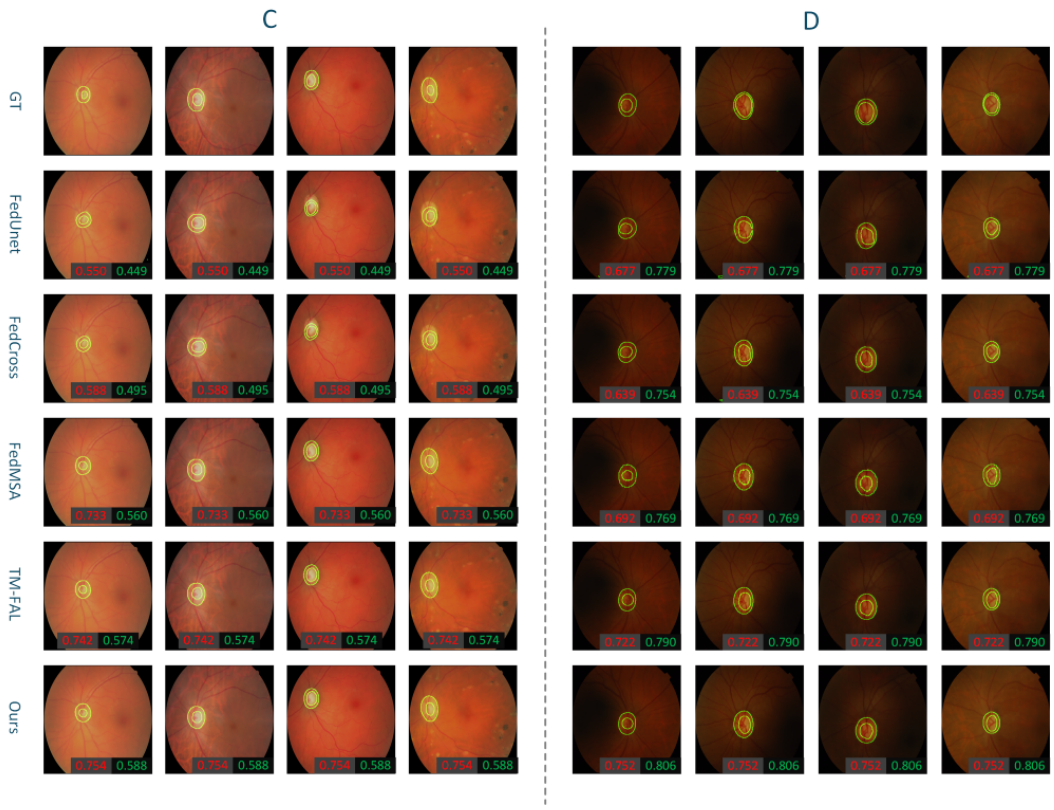

Figure G: Qualitative segmentation comparison on a Fundus image from test clients C and D, highlighting optic disc (OD) and optic cup (OC). From top to bottom: ground truth (GT), FedUnet, FedCross Xu et al. (2024), FedMSA Liu et al. (2024), TM-FAL Yan et al. (2025), and our proposed FedOPS. The Dice values for OD (red) and OC (green) are indicated in the bottom-right corner of each mask.