Supplementary Materials: Personalized Mixture of Experts for Multi-Site Medical Image Segmentation

A. Limitations and Failure Case Analysis

The results show that P-MoLE's overall performance across all sites is better than that of the SOTA models. However, it performs worse than the SOTA on Site C of EndoPolyp dataset. We investigate the reasons behind this poor performance. When all locally trained models fail to identify a target during inference, no useful information is transmitted through the SAM to produce the segmentation. We illustrate this scenario in Figure 1 where all four models completely mis-segment the polyp, leading to predictions that lack relevant information related to the ground truth. We hypothesize that as long as there are at least two good-quality segmentations, P-MoLE can distinguish which predictions to weigh heavily and which to ignore. This hypothesis comes from the ablation study in Tab. **??**, showing a large increase in performance from n=1 to n=2.

In conclusion, in cases where all members of the team of experts produce poor-quality segmentations, P-MoLE makes entirely incorrect predictions. In the future, we plan to solve this by designing better architectures for the individual local models so these misses are avoided.

In federated learning or personalized federated learning, the weights are shared with the centralized server in each round of federation, while in P-MoLE, we share this only once. To this end, we make the same assumption as many federated learning papers that we can not infer the training data from the shared weights [9, 18, 19].

B. Detailed Dataset Descriptions

Endoscopic polyp (EndoPolyp) dataset contains a total of 2187 samples, including images and corresponding masks, which have been divided into four different sites having 1000, 380, 196, 612 samples, respectively, according to work [2, 3, 6, 13]. All images and masks are resized to 384 × 384 and divided into train-test sets where the train set contains 900, 328, 170, and 550 samples, respectively, and the rest are considered as the test set according to work [5].

Retinal Fundus (RIF) dataset [1, 12, 14] comprises 1060 images and corresponding masks, which are divided into four different sites as per the work done by [8]. Each site contains 101, 159, 400, and 400 samples. All images in the dataset have been processed according to the work done by [5], where they are resized from their original size of 800 × 800 to 384 × 384 by performing center-cropping. The dataset has been split into a train-test set following the work done by [8], with the train set consisting of 80, 129, 320, and 320 samples for each site, respectively, while the remaining samples are included in the test set.

Table 1. Quantitative performance on RIF dataset. We present the performance on each site and overall performance by taking their mean. The best performance is highlighted in bold. This table shows the results when one element in the ensemble is intentionally poisoned. Results show that even with this poisoned model, P-MoLE can still achieve state of the art performance.

Site	Dice ↑				
	A	B	C	D	Average
Local Models	93.92	88.36	91.03	91.20	91.13
FedAvg [10]	86.86	77.72	87.17	88.28	85.01
FineTune [17]	92.19	89.91	91.77	92.21	91.52
DITTO [7]	92.02	90.34	91.78	92.00	91.53
FedRep [4]	92.23	89.41	91.71	92.19	91.38
FedBABU [11]	92.53	89.20	91.80	92.67	91.55
LC-Fed [16]	92.63	90.62	92.39	92.91	92.14
FedDP [15]	92.96	91.33	92.46	93.03	*92.44*
P-MoLE (Poisoned)	94.02	91.21	92.26	92.81	92.56
P-MoLE (ours)	**95.33**	**92.66**	**94.01**	**94.03**	**94.01**

C. Impact of Bad Local Models

We conducted an ablation study to investigate whether the performance of P-MoLE depends on the good training of the local models. In other words, whether a bad local model can significantly degrade the performance of P-MoLE. To validate this, we run a simple experiment where one of the frozen models in the ensemble is intentionally poisoned by totally randomizing all weights. Thus, this model will only predict garbage. Then, P-MoLE is trained with this poisoned model and included in the team of experts. Results, shown for RIF in Tab 1, outline that even with one poisoned model, while slightly worse than without poisoned models, the poisoned P-MoLE (92.56 Dice) still performs slightly better than the state of the art (FedDP,

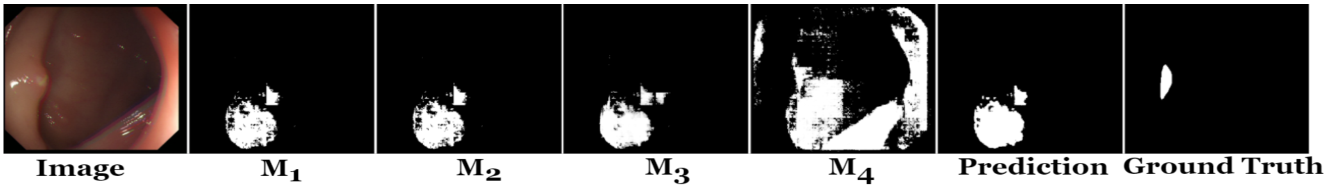

Figure 1. Demonstration of a case where P-MoLE fails to make a correct prediction. Here, M_1, M_2, M_3, and M_4 denote the predictions from four local models, respectively. When these noisy predictions pass through the P-MoLE, it makes completely inaccurate predictions as no relevant feature is present.

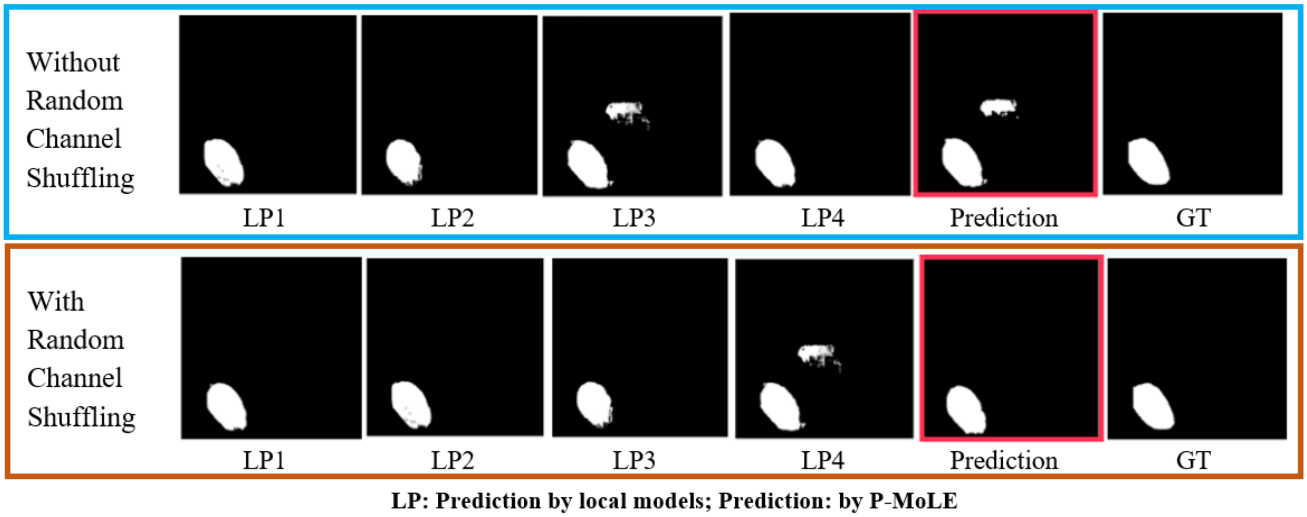
Without Random Channel Shuffling
LP1
LP2
LP3
LP4
Prediction
GT
With Random Channel Shuffling
LP1
LP2
LP3
LP4
Prediction
GT
LP: Prediction by local models; Prediction: by P-MoLE

Figure 2. Example from Site C in the EndoPolyp dataset showing the effect of random channel shuffling. In the top example, we see P-MoLE has overfit to LP3, copying it almost identically for the final prediction. Using random shuffling, LP3 will not be in the same location every time, thus the model cannot hyper focus on one channel, forcing it to pay attention to all channels. The result is that the noise is ignored and a much better segmentation.

92.44 Dice) with one less site's data and a model actively predicting junk. Each model is trained 5 times to ensure statistical relevance. This shows the resiliency of the algorithm against poor-quality inferences and its ability to recognize poor models and ignore them in the final prediction.

D. Impact of Random Channel Shuffling

Random channel/prediction shuffling only applies to the training of P-MoLE and not to inference or the training of the local models. This is just referring to the channel order within the team of experts and does not affect the training of the local models which are trained independently. A qualitative example is shown in Fig. 2. In this case, instead of overfitting to Local Prediction 3 (LP3) and copying it exactly, shuffling these channels around the model ignores the noise in LP3 and produces a better segmentation.

References

[1] Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis and Stereology*, 39(3):161–167, 2020. 1

[2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 1

[3] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. 1

[4] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021. 1

[5] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmen-

tation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 1

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020. 1

[7] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021. 1

[8] 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*, pages 1013–1023, 2021. 1

[9] Ming Y Lu, Richard J Chen, Dehan Kong, Jana Lipkova, Rajendra Singh, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Federated learning for computational pathology on gigapixel whole slide images. *Medical image analysis*, 76:102298, 2022. 1

[10] 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*, pages 1273–1282. PMLR, 2017. 1

[11] Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021. 1

[12] 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. 1

[13] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014. 1

[14] 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. 1

[15] Jiacheng Wang, Yueming Jin, Danail Stoyanov, and Liansheng Wang. Feddp: Dual personalization in federated medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023. 1

[16] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pages 456–472. Springer, 2022. 1

[17] Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019. 1

[18] Jeffry Wicaksana, Zengqiang Yan, Dong Zhang, Xijie Huang, Huimin Wu, Xin Yang, and Kwang-Ting Cheng. Fedmix: Mixed supervised federated learning for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 1

[19] Rui Yan, Liangqiong Qu, Qingyue Wei, Shih-Cheng Huang, Liyue Shen, Daniel L Rubin, Lei Xing, and Yuyin Zhou. Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging. *IEEE Transactions on Medical Imaging*, 42(7):1932–1943, 2023. 1