Federated Modality-Specific Encoders and Multimodal Anchors
for Personalized Brain Tumor Segmentation

Qian Dai1,*, Dong Wei2,*, Hong Liu3,*, Jinghan Sun2,3, Liansheng Wang1,†, Yefeng Zheng2

1School of Informatics, Xiamen University, Xiamen, China
2Jarvis Research Center, Tencent Youtu Lab / Tencent Healthcare (Shenzhen) Co., Ltd., Shenzhen, China
3School of Medicine, Xiamen University, Xiamen, China
{daiqian,liuhong,jhsun}@stu.xmu.edu.cn, lswang@xmu.edu.cn, {donwei,yefengzheng}@tencent.com

Abstract

Most existing federated learning (FL) methods for medical image analysis only considered intramodal heterogeneity, limiting their applicability to multimodal imaging applications. In practice, it is not uncommon that some FL participants only possess a subset of the complete imaging modalities, posing inter-modal heterogeneity as a challenge to effectively training a global model on all participants’ data. In addition, each participant would expect to obtain a personalized model tailored for its local data characteristics from the FL in such a scenario. In this work, we propose a new FL framework with federated modality-specific encoders and multimodal anchors (FedMEMA) to simultaneously address the two concurrent issues. Above all, FedMEMA employs an exclusive encoder for each modality to account for the inter-modal heterogeneity in the first place. In the meantime, while the encoders are shared by the participants, the decoders are personalized to meet individual needs. Specifically, a server with full-modal data employs a fusion decoder to aggregate and fuse representations from all modality-specific encoders, thus bridging the modalities to optimize the encoders via backpropagation reversely. Meanwhile, multiple anchors are extracted from the fused multimodal representations and distributed to the clients in addition to the encoder parameters. On the other end, the clients with incomplete modalities calibrate their missing-modal representations toward the global full-modal anchors via scaled dot-product cross-attention, making up the information loss due to absent modalities while adapting the representations of present ones.

Introduction

Federated learning (FL) enables participants to collaboratively train a global model on their collective data without breaching privacy (Li et al. 2020). The decentralized mechanism makes it particularly suitable for privacy-sensitive application scenarios such as medical image analysis (Kaiissis et al. 2020; Adnan et al. 2022; Bercea et al. 2021; Yan et al. 2021, 2020). However, most FL methods for medical image analysis only considered intramodal heterogeneity, limiting their applicability to multimodal imaging in practice.

One such application is brain tumor segmentation in multi-parametric magnetic resonance imaging (MRI) (Iv et al. 2018). Specifically, four MRI modalities (in this work, we refer to MRI sequences as modalities following literature (Doren et al. 2019; Menze et al. 2014; Shen and Gao 2019; Zhou et al. 2021)) are commonly used to provide complementary information and support sub-region analysis: T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2-weighted (T2), and T2 fluid attenuation inversion recovery (FLAIR), where the first two highlight tumor core and the last two highlight peritumoral edema (Fig. 1(a)). When applying FL to such multimodal applications in practice, it is not uncommon that some participant institutes only possess a subset of the full modalities due to different protocols practiced, presenting a new challenge with the inter-modal heterogeneity (Fig. 1(b)) across the participants of FL. In such a scenario, there can be two objectives for FL: 1) collectively training an optimal global model for full-modal input, and 2) obtaining a personalized model for each participant (Chen and Zhang 2022; Wang et al. 2019), adapted for its data characteristics, and more importantly, better than trained locally without FL. To our knowledge, these two objectives were rarely considered together in FL for medical image analysis.

In this paper, we propose a new FL framework with...
federated modality-specific encoders and multimodal anchors (FedMEMA) for brain tumor segmentation. Above all, to handle the distinctively heterogeneous MRI modalities, FedMEMA employs an exclusive encoder for each modality to allow a great extent of parameter specialization. In the meantime, while the encoders are shared between the server and clients, the decoders are personalized to cater to individual participants. Specifically, a multimodal fusion decoder on the server (i.e., a participant with full-modal data) aggregates and fuses representations from the encoders to bridge the distribution gaps between modalities and reversely optimizes the encoders via backpropagation. Meanwhile, multiple anchors are extracted from the fused multimodal representations and distributed to the clients in addition to the encoder parameters. On the other end, the clients with incomplete modalities calibrate their local missing-modal representations toward the global full-modal anchors via the scaled dot-product attention mechanism (Vaswani et al. 2017) to make up the information loss due to absent modalities and adapt representations of present ones. To this end, we simultaneously obtain an optimal server model (for full-modal input) and personalized client models (for specific missing-modal input) from FL without sharing privacy-sensitive information.

In summary, our contributions are as follows:

- We bring forward the inter-modal heterogeneity problem due to missing modalities in FL for medical image analysis and aim to obtain an optimal full-modal server model and personalized missing-modal client models simultaneously with a novel framework coined FedMEMA.
- To tackle the inter-modal heterogeneity, we propose to employ a federated encoder exclusive for each modality followed by a server-end multimodal fusion decoder. Meanwhile, personalized decoders are employed for the clients to allow simultaneous personalization.
- In addition, we propose to extract and distribute multimodal representations from the server to the clients for local calibration of modality-specific features.
- Last but not least, we further enhance the calibration with multi-anchor representations.

Experimental results on the public BraTS 2020 benchmark show that our method achieves superior performance for both the server and client models to existing FL methods and that its novel designs are effective.

**Related Work**

**Brain Tumor Segmentation with Multimodal MRI:** Multimodal MRI is the current standard of care for clinical imaging of brain tumors (Iv et al. 2018). Segmentation and associated volume quantification of heterogeneous histological sub-regions are valuable to the diagnosis/prognosis, therapy planning, and follow-up of brain tumors (Menze et al. 2014). In recent years, deep neural networks (DNNs) significantly advanced state-of-the-art of brain tumor segmentation with multimodal MRI (Chen et al. 2020; Chen, Ding, and Liu 2019; Ding et al. 2020; Myronenko 2018; Zhou et al. 2020). However, these methods were optimized for ideal scenarios where the complete set of modalities was present. In practice, scenarios of missing one or more modalities commonly occur due to image corruption, artifacts, acquisition protocols, allergy to contrast agents, or cost. Therefore, many efforts have been made to accommodate the practical scenarios of missing modalities (Hu et al. 2020; Wang et al. 2021; Ding, Yu, and Yang 2021; Azad, Khosravi, and Merhof 2022). These methods successfully improved DNNs’ feature representation capability against missing modalities—however, only in the centralized setting, limiting their efficacy in privacy-sensitive settings. In this work, we aim to address the missing-modal problem in the FL setting and eliminate the privacy issue.

**Multimodal FL with Data Heterogeneity:** Data heterogeneity is a primary challenge in FL (McMahan et al. 2017). Personalized FL (Tan et al. 2022a) proposed to adapt the global model locally on clients’ data to address this issue. However, it did not consider the heterogeneity due to multimodal data. We are aware of several works for multimodal FL in the natural image domain. Xiong et al. (2022) proposed a co-attention mechanism to fuse the complementary information of different modalities, yet requiring all clients to have access to the same set of modalities. FedIoT (Zhao, Barnaghi, and Haddadi 2022) employed cross-modal autoencoders to learn multimodal representations in an unsupervised manner. However, both methods (Xiong et al. 2022; Zhao, Barnaghi, and Haddadi 2022) only obtained a single global classifier without catering to the personalized needs of modal-heterogeneous clients. Yu et al. (2023) proposed a cross-modal contrastive representation ensemble between the server and modal-heterogeneous clients by sharing a multimodal dataset, which may be unacceptable in strict privacy restrictions like healthcare. In contrast, our framework optimizes a global model for full-modal input and simultaneously customizes a personalized model for each client’s hetero-modal input. It also maintains FL’s data privacy by transmitting population-wise abstracted prototypes instead of image-wise features.

In medical image analysis, heterogeneity issues in multimodal FL have yet to be thoroughly discussed. FedNorm (Bernecker et al. 2022) adapted the normalization parameters for different modalities while sharing common backbone parameters for CT- and MRI-based liver segmentation. Yet, our experiments suggest that merely specializing in normalization parameters is insufficient to deal with the inter-modal heterogeneity in multimodal brain tumor segmentation. In this work, we propose to handle the heterogeneity with modality-specific encoders to allow a greater extent of parameter specialization, followed by a multimodal fusion decoder to aggregate and fuse representations from the encoders and bridge the inter-modal distribution gaps.

**Method**

**Problem Definition:** Let us denote the full set of modalities by $M=\{T1, T1c, T2, FLAIR\}$ and a full-modal input by $X_M \in \mathbb{R}^{M\times D\times H\times W}$, where $D$, $H$, and $W$ are the depth, height, and width of the volume, respectively. We consider a heterogeneous FL setting where a server with
Figure 2: Overview of the proposed FedMEMA framework. FedMEMA employs a federated encoder exclusive for each modality followed by a server-end multimodal fusion decoder. Meanwhile, personalized decoders are used for the clients to allow simultaneous personalization. In addition, multi-anchor multimodal representations are extracted from the server and distributed to the clients for localized adaptive calibration of modality-specific features via cross-attention. ED: edema, ET: enhancing tumor, NET: necrotic and non-enhancing tumor core, and BG: background.

access to a set of full-modal data \(\{(X_M, Y)\}\) coordinates several clients with access to data of incomplete modalities ("missing-modal"), where \(Y \in \mathbb{R}^{N_c \times D \times H \times W}\) is the segmentation mask and \(N_c\) is the number of target classes. In practice, the server may be a major regional hospital, whereas the clients may be smaller local health units. In this work, we mainly address the extreme and most challenging case of monomodal clients, i.e., each client houses data of a specific modality denoted by \(\{(X_m, Y)\}\), where \(X_m \in \mathbb{R}^{1 \times D \times H \times W}\), for a pioneer and exploratory methodology development (some preliminary results in the more common settings of multimodal clients are provided in the supplementary material). Furthermore, we mainly consider the horizontal FL setting where 1) each client has different patients, or 2) a client may share some patients with another, i.e., various clients may have different modalities of the same patients. The latter scenario is practical as a patient may have different imaging studies at different hospitals within a district. However, considering the privacy issue, the clients do not know whether or which of their data overlaps with others.

This work aims to train, via FL, not only an optimal global model that works well with full-modal data but also optimal personalized models that work well in specific missing-modal situations for the clients. The latter is practically meaningful as it is usually difficult for a local health unit to upgrade its imaging protocol quickly, yet still look forward to training an optimal model for its current protocol by participating in the FL.

Framework Overview: Without loss of generality, we use four clients, each with data of a mutually different modality, for method description. As shown in Fig. 2, the server has four modality-specific encoders (one for each modality) and a modal fusion decoder, whose fused features are clustered to produce multimodal anchors. Meanwhile, each client has a modality-specific encoder for local data modality and a private decoder. Additionally, a localized adaptive calibration via cross-attention (LACCA) module calibrates the clients’ missing-modal representations toward the server’s multimodal anchors.

FL with Modality-specific Encoders: In classical FedAvg (McMahan et al. 2017), the server and clients usually share the same network architecture, where the server aggregates and averages the network parameters of the clients and then distributes the averaged parameters back to the clients in a straightforward manner. However, due to the high heterogeneity among the multimodal MRI data (see Fig. 1), our problem setting becomes challenging for this paradigm. Instead, we propose federated modality-specific encoders to handle the distinctively heterogeneous imaging modalities. On the one hand, we adopt an architecture with late fusion strategy (Ding, Yu, and Yang 2021) to compose the global model on the server, including a modality-specific encoder \(E_m\) (parameterized by \(W^m\)) for each modality, a fusion decoder \(D_M\) for multimodal feature aggregation and fusion, and a regularizer (not shown in Fig. 2 for simplicity). The regularizer is a straightforward auxiliary segmentation decoder shared by all modality-specific encoders. It regularizes the encoders to learn the same discriminative features by forcing them to share the decoder parameters. Please refer to (Ding, Yu, and Yang 2021) for details. Given a full-modal

\(\text{Note that our framework is model-agnostic and can be implemented with various non-FL multimodal segmentation models consisting of modality-specific encoders and modal fusion decoder(s) (e.g., Dorent et al. 2019; Shen and Gao 2019; Zhou et al. 2021). In this work, we use (Ding, Yu, and Yang 2021) for demonstration due}\)
input $X_M$, each $E_m$ first extracts features from the corresponding modality $X_m$, followed by $D_M$ fusing multimodal features and generating segmentation masks. On the other hand, each single-modal client has a federated modality-specific encoder $E_m$ and a personalized decoder $D_m$, $E_m$ on the clients shares the same architecture as the server.

In each round of FL, the clients first receive parameters $W^a_m$ from the server to replace its local copy $W^i_m$, where $i \in \{1, \ldots, N_m\}$ and $N_m$ is the number of clients with data of modality $m$, train for $N_e$ epochs on local data, and then send updated $W^i_m$ back to the server. After receiving $W^i_m$, the server averages $W^i_m$ of the same modality (if $N_m > 1$): $W^s_m = \frac{1}{N_m} \sum_i W^i_m$, train for $N_e$ epochs on the full-modal data, and sends updated $W^s_m$ to the clients for the next round. Thus, the server bridges the distribution gaps between modalities with the fusion decoder $D_M$ and utilizes the complementary multimodal information to train each modality-specific encoder $E_m$ via backpropagation.

**Multi-Anchor Multimodal Representation:** Besides aligning modality-specific encoders with the multimodal fusion decoder, the server also generates multi-anchor multimodal representations for the classes of interest, which will be distributed to the clients in addition to the encoder parameters. Liu et al. (2020) proposed communicating encoded representations, which may breach the privacy restriction. On the contrary, some works proposed to transmit category prototypes (Mu et al. 2023; Tan et al. 2022b). Yet a single prototype was highly compressed and may not carry enough representative information for a class, especially considering the significant inter-subject variations in 3D multimodal medical images.

In this work, we propose to extract multiple prototypes (Cui et al. 2020) from the fused multimodal features for each class of interest for enhanced representation power, which we refer to as anchors for their calibration purpose (Ning et al. 2021). Concretely, we extract per-class features from the fused multimodal feature maps of the decoder $D_M$ by masked average pooling using the ground truth mask and apply the K-means method (MacQueen et al. 1967) to the extracted features to obtain $N_c$ anchors. Intuitively, these $N_c$ anchors are the modes of each class’s multimodal distribution. For the $l$th feature scale level, where $l \in \{1, \ldots, 4\}$ for the networks we use, the anchors for all the $N_c$ classes can be collectively denoted by $A_l \in \mathbb{R}^{N_c N_l \times C_l}$, where $C_l$ is the number of feature channels. Empirically, we determine the cluster membership using the most abstract feature level, i.e., $l = 4$, and apply the membership to compute $N_k = 3$ anchors for all levels (see corresponding experiments in the next section). This strategy can well preserve the full-modal information of each class while incurring little network transmission burden. It should also be noted that the few to its outstanding performance and straightforward architecture.

We are aware of strategies for dynamic weight assignment for parameter aggregation based on clients’ data sizes (e.g., Hsu, Qi, and Brown 2020). Although this work focuses on inter-modal heterogeneity and assigns the same amount of data to all clients, we expect the incorporation of dynamic weight assignment to make our method more robust in practice.

Algorithm 1: FedMEMA algorithm. Note that the personalized (non-federated) decoder parameters are not shown below for simplicity.

**Require:** the modality set $M = \{T1, T1c, T2, FLAIR\}$ indexed by $m$, a full-modal training dataset $D_M$ on the server, the number of clients $N_m$ with data of modality $m$, the monomodal training set $D^m$ on client $i$ with data of modality $m$, the number of communication rounds $N_r$, and the number of training epochs $N_e$ in each round.

**Output:** the collection of parameters $W^s_{(m)} = \{W^s_m\}$ for the modality-specific encoders, and the collection of multimodal anchors $A_{(l)} = \{A_l\}$ for different feature scale levels $l$ (plus parameters of the personalized decoders).

**Server executes:**
1: Server executes:
2: Initialize $W^s_{(m)}$, and update $W^s_{(m)}$ by training on $D_M$ for $N_e$ epochs
3: Initialize $A_{(l)}$ by K-means
4: for round $r = 1$ to $N_r$, do
5: for $m \in M$ do
6: for each client $i \in N_m$ do
7: $W^i_m \leftarrow$ ClientUpdate($m, i, W^a_m, A_{(l)}$) \> run on client $i$
8: $W^s_m = \frac{1}{N_m} \sum_i W^i_m$ \> aggregate parameters for modality-specific encoder
9: Update $W^s_{(m)}$ by training on $D_M$ for $N_e$ epochs
10: Update $A_{(l)}$ by exponential moving average
11: for modal $m$ do
12: $W^i_m \leftarrow W^s_m$
13: Update $W^i_m$ by training on $D^m$, with LACCA (Eq. (1)) for $N_e$ epochs
14: return $W^i_m$

class-wise anchors are abstracted from the entire training population on the server, thus carrying little privacy information concerning individuals. To avoid the collapse of the training process due to jumps in cluster centroids as a result of re-clustering at each round (Xie, Girshick, and Farhadi 2016), we treat the anchors as a memory bank and update them smoothly via exponential moving average (EMA; Tarvainen and Valpola 2017): $\bar{a}_c = \omega a_c + (1 - \omega) \bar{a}_c$, where $a_c$ is an anchor for class $c$ in the memory bank and updated by the closest cluster centroid $\bar{a}_c$ of the same class, and $\omega$ is set to 0.999 following Tarvainen and Valpola (2017).

**Localized Adaptive Calibration via Cross-Attention:** In each federated round, the clients receive from the server the multimodal anchors $A_l$, which are used to calibrate local missing-modal representations. Concretely, denoting the final feature map at the $l$th scale level of the encoder by $F_l \in \mathbb{R}^{D \times H \times W \times C_l}$, we reshape $F_l$ to the dimension $D^{HW} \times C_l$. Then, inspired by the attention operation in the Transformer architecture (Vaswani et al. 2017), we treat the reshaped $F_l$ as queries and the multimodal anchors as the keys and values, and calibrate the local representations toward the global multimodal anchors by the cross-attention:

$$F_l^{cal} = \text{Attn}(F_l, A_l) = \text{softmax} \left( \frac{F_l A_l^T}{\sqrt{C_l}} \right) A_l.$$  (1)
Finally, the calibrated features $F^{cal}_k$ are reshaped back and element-wise added to the final features of the same scales of the decoder to participate in the subsequent forward propagation. The calibration process is localized and self-adaptive in that each client locally emphasizes the part of the global multi-anchor multimodal representations that best suits its own data modality and distributions—via the dot-product attention—to yield more powerful models tailored for itself.

To this end, we name it the localized adaptive calibration via cross-attention (LACCA) module. The LACCA module is inserted in all four feature scales of our backbone networks. Note that the multimodal anchors are learnable parameters during FL and are directly used by the clients for inference after training. The complete algorithm of our method is detailed in Algorithm 1.

## Experiments and Results

### Dataset and Experimental Settings

We conduct experiments on the multimodal Brain Tumor Segmentation (BraTS) 2020 dataset (Menze et al. 2014; Bakas et al. 2018), which consists of 369 multi-contrast MRI scans with four sequences: T1, T1c, T2, and FLAIR. The goal is to segment three nested subregions of brain tumors: whole tumor, tumor core, and enhancing tumor. Following Ding, Yu, and Yang (2021), we divide the dataset into 219, 50, and 100 subjects for training, validation, and testing, respectively. The test set is used only for the final model evaluation, whereas the validation set is used for model optimization. Without loss of generality, we design two experimental settings per our problem definition. In setting 1, the training set is evenly divided among the server and four clients at random, i.e., no subject overlap between clients. Thus, the server and clients include 43, 44, 44, and 44 training subjects, respectively. In setting 2, the server data remain unchanged, whereas the rest of the training data are randomly divided into 4 + 1 equal parts where the additional “1” is the common data for all clients. The server and clients include 43, 71, 71, 71, and 71 training subjects, respectively. Meanwhile, the server and all clients in both settings use the same validation and test sets. Note that each client can only access the specific modalities of its assigned data. The mean of the Dice similarity coefficients (mDSC) of the three tumor subregions is employed as the evaluation metric, and the Wilcoxon signed rank test is used to analyze statistical significance.

### Implementation

The proposed FL framework is implemented using PyTorch (1.13.0) and trained with five RTX 2080Ti GPUs, with the server on one GPU and the clients evenly distributed on the rest. We use the RFNet (Ding, Yu, and Yang 2021) as our server network, and its modality-specific encoder and regularizing decoder as the clients’ encoder and decoder, respectively. The LACCA module is implemented with eight attention heads. The input crop size is $80 \times 80 \times 80$ voxels, and the batch size is set to 1 and 3 for the server and clients, respectively. Other settings are the same for the server and clients. The commonly used Dice loss (Milletari, Navab, and Ahmadi 2016) plus the cross entropy loss for medical image segmentation are employed. The Adam optimizer, with its learning rate and weight decay set to 0.0002 and $10^{-5}$, respectively, is leveraged for optimization. We train the networks for 1000 rounds, and in each federated round, the server and clients are trained for one epoch. We follow Ding, Yu, and Yang (2021) for data preprocessing and augmentation. Our code is available at https://github.com/QDaiing/FedMEMA.

### Validating Designs for Multi-Anchor Multimodal Representations

Based on the validation data, we first determine 1) the optimal number of multimodal anchors per class ($N_k$) and 2) the optimal feature scale level $l$ based on which the cluster membership is determined. We fix either of them to reduce the search space while varying the other. The results are shown in Table 1. Although the results look fairly stable, we select $N_k = 3$ and $l = 4$ for evaluation and comparison with other methods on the test data due to their highest performances in both the clients’ average and the server’s mDSCs compared with alternative values. We conjecture that $l = 4$ (i.e., the most abstract feature level) works the best due to its great capability of abstraction and denoising despite the relatively low resolution.

<table>
<thead>
<tr>
<th>$N_k$</th>
<th>FLAIR</th>
<th>T1c</th>
<th>T1</th>
<th>T2</th>
<th>Avg</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60.45*</td>
<td>75.34*</td>
<td>54.73</td>
<td>57.12*</td>
<td>61.91*</td>
<td>83.80*</td>
</tr>
<tr>
<td>2</td>
<td>60.22*</td>
<td>77.29*</td>
<td>57.66*</td>
<td>59.08*</td>
<td>63.56*</td>
<td>83.91*</td>
</tr>
<tr>
<td>3</td>
<td><strong>62.52</strong></td>
<td><strong>76.37</strong></td>
<td><strong>57.26</strong></td>
<td><strong>60.36</strong></td>
<td><strong>64.13</strong></td>
<td><strong>84.17</strong></td>
</tr>
<tr>
<td>4</td>
<td>61.16*</td>
<td>76.71*</td>
<td>56.28*</td>
<td>59.67*</td>
<td>63.45*</td>
<td>83.85*</td>
</tr>
<tr>
<td>5</td>
<td>59.91*</td>
<td><strong>77.37</strong></td>
<td>57.21*</td>
<td>59.05*</td>
<td>63.38*</td>
<td>83.95*</td>
</tr>
<tr>
<td>6</td>
<td>70.93*</td>
<td>76.69*</td>
<td><strong>57.84</strong></td>
<td>58.93*</td>
<td>63.60*</td>
<td><strong>83.46</strong></td>
</tr>
</tbody>
</table>

Table 1: Results of experimental setting 1 on the validation set in mDSC (%). Top: varying $N_k$ (number of multimodal anchors per class) with $l = 4$. Bottom: varying the feature scale level $l$ (with $N_k = 3$) of the multimodal fusion decoder $D_M$, on which the cluster membership is determined; $l = 4$ indicates the most abstract level of the smallest scale (i.e., at the bottleneck between the encoders and decoder), and “1–4” concatenates features of all four levels together for clustering. FLAIR, T1c, T1, and T2 indicate performance of the clients with the corresponding data modalities, Avg indicates their average, and “$*$” indicates server performance. $*: p < 0.05$ comparing against $N_k = 3$ (top) and $l = 4$ (bottom), respectively, in each column.

3In this paper, we primarily focus on the inter-modal heterogeneity due to missing modalities but ignore the intramodal heterogeneity due to institutions, i.e., potential distribution discrepancies between data of the same modality but from different institutions.
be noted that as CreamFL requires sharing the server data and clients for the counterparts in other methods. It should use the same networks) and the same networks as our server derivatives (in FedAvg infrastructure, the server and clients use the same networks as our clients' for FedAvg and dIoT's autoencoding clients to supervised networks. Also, so on the server data as our method, not originally conduct server-end training, we make them do its derived methods (i.e., PerFL and FedNorm), which did. 

tions are made to ensure fair comparison: 1) for FedAvg and the methodological principles unchanged, necessary adapta-

ized models, including FedNorm (Bernecker et al. 2022), FedMSplit (Chen and Zhang 2022), CreamFL (Yu et al. 2023), FedIoT (Zhao, Barnaghi, and Haddadi 2022), and PerFL (Wang et al. 2019). Under the premise of keeping

... classFedAvg (McMahan et al. 2017) and several up-
to-date approaches to FL on multimodal data or personal-
ized models, including FedNorm (Bernecker et al. 2022), FedMSplit (Chen and Zhang 2022), CreamFL (Yu et al. 2023), FedIoT (Zhao, Barnaghi, and Haddadi 2022), and PerFL (Wang et al. 2019). Under the premise of keeping the methodological principles unchanged, necessary adaptations are made to ensure fair comparison: 1) for FedAvg and its derived methods (i.e., PerFL and FedNorm), which did not originally conduct server-end training, we make them do so on the server data as our method, and 2) we change FedIoT's autoencoding clients to supervised networks. Also, we use the same networks as our clients' for FedAvg and derivatives (in FedAvg infrastructure, the server and clients use the same networks) and the same networks as our server and clients for the counterparts in other methods. It should be noted that as CreamFL requires sharing the server data with all clients, it violates the privacy restriction in the medical context and increases the training data for the clients. Lastly, as RFNet (Ding, Yu, and Yang 2021) was originally designed for both full- and missing-modal segmentation after training with full-modal data, we also train it on the server data in its original recipe and evaluate its performance for comparison.

The results are shown in Table 2. As the comparative trends are similar in both settings, we mainly describe setting 1 below. Without modality-specific parameters, FedAvg and PerFL mostly yield worse performance than the baseline local models. This indicates that the inter-modal heterogeneity impedes the classical FL from effectively utilizing the extra data on the clients. By specializing the normalization parameters for different modalities, FedNorm achieves slight improvements in average mDSC across the clients but with slight to modest decreases in the server’s performance. The more complex and advanced CreamFL, FedMSplit, and FedIoT frameworks can achieve comprehensive improvements over the baseline; especially, FedIoT exceeds...  

<table>
<thead>
<tr>
<th>Method</th>
<th>Setting 1</th>
<th>FLAIR</th>
<th>T1c</th>
<th>T1</th>
<th>T2</th>
<th>Avg</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local models</td>
<td></td>
<td>FLAIR</td>
<td>T1c</td>
<td>T1</td>
<td>T2</td>
<td>Avg</td>
<td>S</td>
</tr>
</tbody>
</table>
| RFNet        |           | 55.21 | 64.60 | 42.98 | 54.84 | 54.41 | 79.91 |*
| FedAvg       |           | 54.04 | 62.40 | 36.60 | 54.11 | 51.78 | 78.42 |*
| PerFL        |           | 54.94 | 63.56 | 39.08 | 53.57 | 52.78 | 78.71 |*
| FedNorm      |           | 57.06 | 68.45 | 39.90 | 56.16 | 55.39 | 77.72 |*
| CreamFL      |           | 55.07 | 66.24 | 47.01 | 56.92 | 56.31 | 80.30 |*
| FedMSplit    |           | 56.54 | 69.58 | 46.31 | 59.96 | 58.09 | 81.01 |*
| FedIoT       |           | 58.85 | 71.14 | 49.96 | 61.07 | 60.25 | 80.23 |*
| FedMEMA (ours) |       | 61.02 | 71.91 | 53.56 | 61.55 | 62.01 | 82.11 |*

<table>
<thead>
<tr>
<th>Method</th>
<th>Setting 2</th>
<th>FLAIR</th>
<th>T1c</th>
<th>T1</th>
<th>T2</th>
<th>Avg</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local models</td>
<td></td>
<td>FLAIR</td>
<td>T1c</td>
<td>T1</td>
<td>T2</td>
<td>Avg</td>
<td>S</td>
</tr>
</tbody>
</table>
| RFNet        |           | 55.35 | 69.67 | 44.24 | 57.47 | 56.68 | 79.91 |*
| FedAvg       |           | 55.59 | 65.47 | 42.65 | 56.61 | 55.08 | 78.19 |*
| PerFL        |           | 56.70 | 64.24 | 45.16 | 57.43 | 55.88 | 80.09 |*
| FedNorm      |           | 55.78 | 70.91 | 50.75 | 51.80 | 57.31 | 78.28 |*
| CreamFL      |           | 59.98 | 69.54 | 50.05 | 59.55 | 59.78 | 81.55 |*
| FedMSplit    |           | 58.87 | 70.70 | 50.41 | 60.11 | 60.02 | 80.86 |*
| FedIoT       |           | 60.47 | 71.96 | 52.49 | 61.03 | 61.48 | 81.47 |*
| FedMEMA (ours) |       | 62.84 | 73.49 | 56.46 | 61.58 | 63.39 | 83.27 |*

Table 2: Experimental results on the test set in mDSC (%). FLAIR, T1c, T1, and T2 indicate the clients’ performance with the corresponding data modalities, Avg indicates their average, and “S” indicates server performance. *: p < 0.05 comparing against our method in each column.

<table>
<thead>
<tr>
<th>Ablation</th>
<th>Server</th>
<th>Federated</th>
<th>LACCA</th>
<th>FLAIR</th>
<th>T1c</th>
<th>T1</th>
<th>T2</th>
<th>Avg</th>
<th>S</th>
</tr>
</thead>
</table>
| (a)     | E&D    | D         | -     | 43.81 | 48.83 | 18.34 | 43.59 | 38.64 | 80.05 |*
| (b)     | E&D    | E         | -     | 51.48 | 58.51 | 36.35 | 47.97 | 48.58 | 81.11 |*
| (c)     | 4E&D   | 4E        | -     | 57.14 | 72.95 | 50.67 | 58.01 | 59.69 | 82.11 |*
| (d)     | 4E&D   | 4E        | Mono-anchor | 60.45 | 75.34 | 54.73 | 57.12 | 61.91 | 83.80 |*
| (e)     | Ours   | 4E&D      | 4E    | 62.52 | 76.37 | 57.26 | 60.36 | 64.13 | 84.17 |*

Table 3: Ablation study on the validation set in experimental setting 1 using mDSC (%). “E” and “D” are encoder and decoder, respectively. FLAIR, T1c, T1, and T2 indicate the clients’ performance with the corresponding data modalities, Avg indicates their average, and “S” indicates server performance. *: p < 0.05 comparing against our method in each column.

Our preliminary experiments empirically showed that with the server-end training, they performed better than without it.
in the clients’ average mDSC by close to 6%. In comparison, our FedMEMA further improves upon FedIoT by \(\sim 2\%\) in both the clients’ average and server mDSCs and achieves the best performance for all clients and the server. We attribute our method’s advantages to the adaptive calibration via cross-attention between local monomodal and global full-modal representations. Remarkably, besides improving the clients’ performance with locally personalized models, FedMEMA also substantially enhances the server’s performance by effectively exploiting the clients’ data of heterogeneous modalities. Meanwhile, the RFNet, despite being a robust model for various missing-modal situations, yields performance comparable to the baseline local models, probably because it does not use the extra data on clients. Lastly, the performance in setting 2 is generally better than that in setting 1, likely due to the more training data available on each client. Fig. 3 shows example segmentation results.

Ablation Study: We conduct thorough ablation studies to validate the efficacy of our novel framework design, including the federated modality-specific encoders and personalized decoders, the LACCA module, and the multi-anchor multimodal representations. The results are presented in Table 3, where the first two rows are variants of the classical FedAvg (McMahan et al. 2017) with either the encoder or decoder federated, and the last three are variants of our proposed method. It can be seen that, among the variants of FedAvg, federating the encoder while personalizing the decoder (row (b)) outperforms the reverse (row (a)). On top of that, our modality-specific encoders (row (c)) achieve substantial performance boosts over the FedAvg family, e.g., the clients’ average mDSC improves by \(\sim 11\%\) compared with row (b), demonstrating the effectiveness of the FL architecture in our problem setting. Row (d) additionally incorporates the LACCA module but with mono-anchor representations obtained by averaging all server data, achieving further improvements, especially on the clients (\(> 2\%\) in average mDSC). This suggests that multimodal representations are effective for the adaptive calibration of local monomodal features. The last row is our full model with multi-anchor representations, yielding the best performances in both average missing-modal mDSC across the clients (\(\sim 64\%\)) and full-modal mDSC on the server (\(\sim 84\%\)). These results indicate that multiple anchors represent the 3D multimodal medical data better than the mono ones, as expected.

Conclusion

In this paper, we proposed FedMEMA—a new FL framework with federated modality-specific encoders and multimodal anchors for brain tumor segmentation in multi-parametric MRI with missing-modal clients, and demonstrated its superior performance to existing FL methods on the public BraTS 2020 benchmark in the extreme case of monomodal clients. In future work, we plan to evaluate FedMEMA with the more common settings of hetero-modal clients and more datasets. In addition, this paper mainly focused on the inter-modal heterogeneity due to missing modalities but ignored the intramodal heterogeneity due to institutions, which was a limitation. We also plan to consider both types of heterogeneities together in the future.
Acknowledgments
This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0109500/2020AAA0109501, and in part by the National Natural Science Foundation of China (Grant No. 62371409).

References
IEEE/CVF International Conference on Computer Vision, 9112–9122.


