# MuCALD-SPLITFED: CAUSAL-LATENT DIFFUSION FOR PRIVACY-PRESERVING MULTI-TASK SPLIT-FEDERATED MEDICAL IMAGE SEGMENTATION

*Chamani Shiranthika, Hadi Hadizadeh, and Parvaneh Saeedi*

School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada

## ABSTRACT

Federated Learning enables decentralized training by aggregating model updates across clients without sharing raw data, while Split Federated Learning further partitions the model between clients and a server to reduce computation and communication at the client side. However, decentralized medical institutions rarely operate on a single shared task, making standard Federated and SplitFed collaborations poorly aligned with real clinical workflows. Multi-task FL extends these frameworks by allowing clients to handle different tasks, but often introduces instability and privacy vulnerabilities. This study proposes **MuCALD-SplitFed**, a multi-task SplitFed framework that integrates causal representation learning and latent diffusion. Experiments show MuCALD-SplitFed consistently improves segmentation, while baseline SplitFed fails to converge. The proposed approach further reduces information leakage at split points, mitigating reconstruction-based and membership inference attacks. Additionally, MuCALD SplitFed outperforms state-of-the-art personalized FL and multi-task FL approaches. The code repository is: `https://github.com/ChamaniS/MuCALD_SplitFed`.

***Index Terms*—** Multi-task SplitFed Learning, Medical Image Segmentation, Causal Latent Diffusion, Privacy-preservation

## 1. INTRODUCTION

Medical image segmentation plays a critical role in clinical workflows. Hospitals, clinics, and research laboratories often differ in imaging modalities, annotation quality, acquisition protocols, population demographics, etc. Practical medical AI systems therefore require learning frameworks that can generalize across heterogeneous tasks while preserving privacy. Decentralized learning paradigms such as Federated Learning (FL) [1], Split Learning (SL) [2], and Split Federated Learning (SplitFed) [3] enable collaborative model training without sharing raw data. However, these approaches implicitly assume all clients solve the same task—the non-IID assumption [1]. which is unrealistic in real clinical settings.

Thanks to Natural Sciences and Engineering Research Council (NSERC) of Canada for funding.

Multi-task learning relaxes this constraint by allowing clients to train on domain-specific tasks while benefiting from shared global representations. In the literature, *multi-task* refer to clients with non-IID institutional data, or clients working on related but different tasks (which is neither IID nor non-IID), or setups where multiple objectives (e.g., segmentation and classification) are learned jointly. Multi-task FL methods- MOCHA [4], sheaf-based FL [5], FedAlign [6], and multi-task SplitFed variants- FedBone [7], FedMSplit [8], homomorphically encrypted multi-task SL [9] demonstrate effective collaboration under heterogeneous tasks.

Multi-task SplitFed systems often introduce convergence instability, as clients optimize heterogeneous objectives across different data distributions and label spaces. During federated aggregation, gradients from different tasks can conflict, leading to unstable or oscillatory updates. The shared latent representation at split points entangles task-specific features, amplifying cross-task interference and hindering convergence. Multi-task SplitFed further introduces generalization and privacy challenges. Split-point communications can leak sensitive information, enabling attacks such as membership inference, reconstruction, property inference, and task-driven client drift. These risks are amplified by diverse task semantics in shared representations.

Addressing these vulnerabilities requires going beyond correlation-based representations. Causality explicitly models cause–effect relationships rather than statistical associations, enabling models to distinguish task-relevant structure from spurious correlations [10]. Recent research in causality in medical AI shows that causal representations improve robustness, reduce bias, and support trustworthy clinical decision-making [11]. Causality-driven models- C-CAM [12], CauSSL[13], and MACAW [14] demonstrate that modelling cause-effect structure yields stronger generalization under distribution shifts. Causal diffusion models— CausalDiffAE [15], Diff-structural causal model (SCM) [16], and latent-causal DDPMs [17] indicate that combining causality and diffusion disentangles semantic factors, suppresses spurious correlations, and enables fine-grained control. However, these ideas have not been explored in multi-task SplitFed.

To address these gaps, we propose MuCALD-SplitFed, a causal, diffusion-driven multi-task SplitFed framework for privacy-preserving medical image segmentation. The contri-

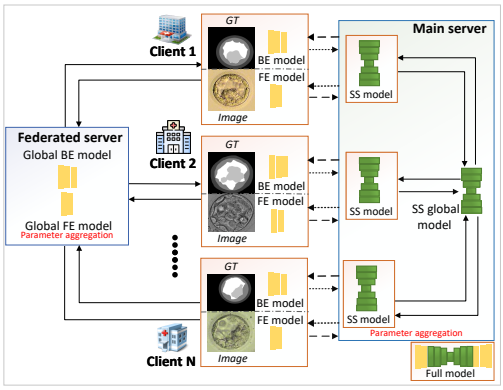



**Fig. 1**. Baseline SplitFed where all clients collaborate on a single shared task (blastocysts segmentation).

butions are: integrating causal modelling within the SplitFed latent space for multi-task segmentation in heterogeneous clinical settings; combining causal representations with latent diffusion and adversarial denoising to improve robustness and reduce information leakage at split points; disentangling task-relevant causal factors from domain-specific information to enhance segmentation performance and privacy; and extensive evaluation on five heterogeneous medical datasets over Baseline SplitFed and state-of-the-art multi-task and personalized FL methods. Section 2 introduces the MuCALD-SplitFed methodology, Section 3 presents experiments and SoTA comparisons, and Section 4 concludes the paper.

## 2. PROPOSED MUCALD-SPLITFED

Our Baseline SplitFed architecture, [18, 19], based on [3] (Fig. 1), partitions a model into front-end (FE), server-side (SS), and back-end (BE) components. Clients receive copies of the global FE and BE models from the federated server and the SS model from the main server. Each client then trains locally for several epochs, after which parameters are aggregated via federated averaging [1] and redistributed. This process repeats for multiple rounds until convergence. Proposed MuCALD-SplitFed framework (left side of Fig. 2) consists of three stages: (1) proxy-label construction and causal graph discovery, (2) causal representation and diffusion module (CRDM), and (3) domain-adversarial causal alignment (DACA). Each stage improves robustness, reduces leakage, or mitigates domain shift, respectively, without added cross-client dependencies.

### 2.1. Proxy-label construction and causal graph discovery

First, domain-specific morphological and intensity features are extracted from the five datasets as proxy labels for weakly supervised causal learning.

- Blastocyst [21]: zona pellucida thickness variation (ZPTV), ZP area, blastocoel (BL) area, blastocyst diameter, BL symmetry, inner cell mass (ICM) area/ solidity/ compactness/ entropy/ brightness, trophectoderm (TE) area/brightness/granularity, embryo expansion grade, bubble count/area, and ZP sharpness.
- FHPsAOPMSB [22]: fetal head size, perimeter, major/ minor axes, circularity, solidity, intensity statistics, bounding-box dimensions, symphysis features, head–symphysis geometric relations, and vertical alignment.
- HAM10K [23], MosMed [24], and Kvasir-SEG [25]: lesion size, perimeter, compactness, bounding-box size, orientation, solidity, asymmetry, intensity statistics, RGB/ HSV statistics, colour variance, and entropy.

Dataset-specific causal graphs are then learned from these proxy features using Notears-MLP [20]. Three strongest causal relations per dataset are selected (right side of Fig. 2).

### 2.2. Causal Representation and Diffusion Module (CRDM)

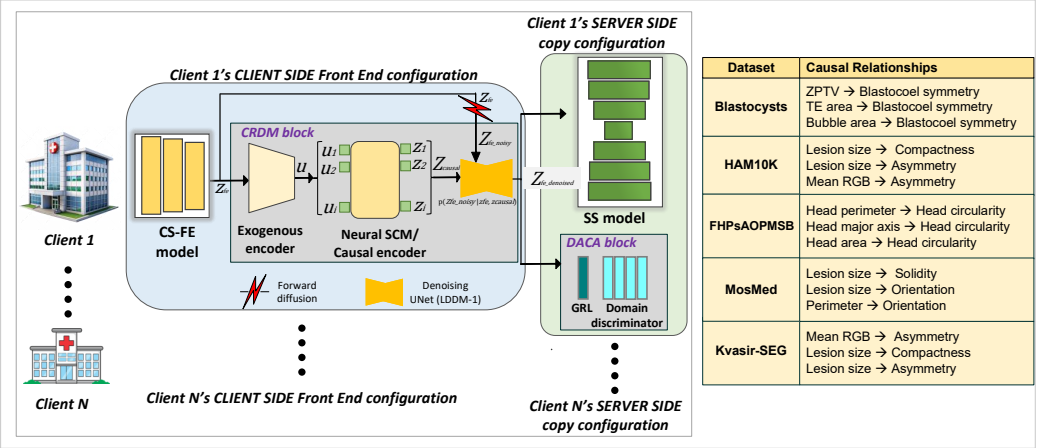
The CRDM block consists of the exogenous encoder, neural SCM, and the denoising LDDM, followed by a forward diffusion step. First the CS-FE model produces feature maps $z_{fe}$, which the exogenous encoder maps to low-dimensional exogenous variables $u = z_{exo} = f_{exo}(z_{fe})$. Second, the Neural-SCM learns a causal factorization that produces causally structured latents ($z_{causal} = f_{SCM}u$). Third, forward diffusion is applied to the encoder features ($z_{fe\_noisy} = forward\_diffusion(z_{fe})$), to obfuscate transmitted activations. Next, a causal diffusion decoder (LDDM-1) denoises the noisy latents $z_{fe\_noisy}$ conditioned on the causal latents $z_{causal}$, generating $z_{denoised}$, which will be sent to the server-side. These causally structured, diffusion-noised latents expose less reconstructive information than standard SplitFed activations, reducing reconstruction and membership inference risks.

### 2.3. Domain-Adversarial Causal Alignment (DACA)

The server-side DACA block enforces domain invariance and suppresses task-specific biases in the received latents. It comprises a gradient reversal layer (GRL) [26], which inverts gradients flowing into the client, and a lightweight domain discriminator, which attempts to predict the originating client or domain. During training, the segmentation network attempts to fool the discriminator, producing domain-agnostic representations. The discriminator penalizes domain-specific leakage, improving privacy and cross-task generalization.

### 2.4. Optimization objective

The Neural-SCM is weakly supervised using proxy labels, yielding proxy-alignment losses $L^{(1)}_{\text{proxy}}$ and $L^{(2)}_{\text{proxy}}$ at the two




| Dataset | Causal Relationships |
|---|---|
| Blastocysts | ZPTV → Blastocoel symmetry<br>TE area → Blastocoel symmetry<br>Bubble area → Blastocoel symmetry |
| HAM10K | Lesion size → Compactness<br>Lesion size → Asymmetry<br>Mean RGB → Asymmetry |
| FHPsAOPMSB | Head perimeter → Head circularity<br>Head major axis → Head circularity<br>Head area → Head circularity |
| MosMed | Lesion size → Solidity<br>Lesion size → Orientation<br>Perimeter → Orientation |
| Kvasir-SEG | Mean RGB → Asymmetry<br>Lesion size → Compactness<br>Lesion size → Asymmetry |

**Fig. 2**. Left: Proposed MuCALD SplitFed (CS-FE to SS split point). Right: Causal relationships discovered using Notears-MLP [20] for Neural SCM modeling. A symmetric configuration operates at the SS to CS-BE split point.

split points. Local model training jointly minimizes proxy-alignment losses, segmentation loss $L_{\text{seg}}$, diffusion regularization losses ($L_{\text{diff}}^{(1)}$ and $L_{\text{diff}}^{(2)}$), KL divergence terms for the exogenous and causal latents ($L_{\text{KL}u}$, $L_{\text{KL}z}$), and the adversarial domain losses ($L_{\text{adv}}^{(1)}$ and $L_{\text{adv}}^{(2)}$) as in Equation 1.

$$L = \lambda_{\text{seg}} L_{\text{seg}} + \lambda_{\text{proxy}} \left( L_{\text{proxy}}^{(1)} + L_{\text{proxy}}^{(2)} \right) + \lambda_{\text{diff}} \left( L_{\text{diff}}^{(1)} + L_{\text{diff}}^{(2)} \right) + \lambda_{\text{KL}u} L_{\text{KL}u} + \lambda_{\text{KL}z} L_{\text{KL}z} + \lambda_{\text{adv}} \left( L_{\text{adv}}^{(1)} + L_{\text{adv}}^{(2)} \right). \quad (1)$$

After each communication round, the client copies of SS models, LDDMs, exogenous encoders, and Neural-SCMs are aggregated. FE and BE remain local, preserving personalization and preventing leakage of task-specific attributes.

## 3. EXPERIMENTAL RESULTS

### 3.1. Datasets and model training

Our multi-task SplitFed framework consists of 5 clients, each occupying a medical imaging dataset. Client 1 uses **Blastocyst** dataset [21] (781 RGB human day-5 embryo images annotated into ZP, TE, BL, ICM, and background). Client 2 holds **HAM10K** dataset [23] (10,015 dermatoscopic RGB images for skin lesion and background segmentation). Client 3 uses **FHPsAOPMSB** dataset [22] (4,000 intrapartum transperineal ultrasound images segmented into fetal head, pubic symphysis, and background). Client 4 uses **MosMed** dataset [24] (2800 lung computed tomography scan images, segmented into the affected area and background). Client 5 holds **Kvasir-SEG** dataset [25] (1,000 endoscopic RGB polyp images segmented into abnormal conditions (lesion/polyp/ulcer) and background). Each client has fixed test sets (70, 1,000, 800, 547, 100), with the remaining data split 85%,/15% for training/validation. The datasets span diverse modalities with significant domain shifts, while causal modeling reduces data dependency, supporting generalization. Augmentations include flips, rotations, RGB shifts, normalization, and brightness/contrast adjustments. Models were trained with soft dice loss [27] and a cosine diffusion schedule. Evaluation used class wise-average intersection over union [28] with and without background (IoU W/B & IoU N/B), precision, recall, F1 score, hausdorff distance (HD95), and average symmetric surface distance (ASSD). We used 24 communication rounds with 5 local epochs per client; 2 warm-up epochs (segmentation only), 3 ramp-up epochs with increasing proxy-label weights, and a final epoch.

### 3.2. Performance comparison

We compare Baseline SplitFed and MuCALD SplitFed with UNet [29], UNet3+ [30], and SwinUNet [31] split models. MuCALD SplitFed consistently achieves superior segmentation and privacy performance (Table 1). Fig. 3 shows unstable IoU N/B for Baseline SplitFed, while MuCALD SplitFed converges stably. Instability from gradient conflicts and entangled representations in Baseline SplitFed is mitigated by causal disentanglement and domain alignment. Further, comparisons are done with SoTA personalized FL approaches (FedPer [32], FedRep [33], FedBN [34], FedProx [35], SCAFFOLD [36]), and SoTA multi-task learning approaches (MOCHA [20], FedEM [37]). Table 2 provides a qualitative comparison of the predictions. Table 3 reports the per-client segmentation and split-point reconstruction performance.

We further performed ablation studies with MS-UNet: (1) CRDM only, (2) DACA only, (3) disabling causal graph discovery, (4) disabling diffusion, and (5) disabling forward noising (Table 4). Results show that all components contribute to performance. Removing diffusion causes the largest degradation, highlighting its importance for stable and robust

| Method | Dice | IoU W/B | IoU N/B | Preci -sion | F1 Recall | F1 Score | HD95 | ASSD |
|---|---|---|---|---|---|---|---|---|
| BS-UNet | 0.394 | 0.367 | 0.045 | 0.547 | 0.435 | 0.415 | 78.911 | 31.018 |
| **MS-UNet** | **0.816** | **0.712** | **0.579** | **0.814** | **0.828** | **0.816** | **37.249** | **6.203** |
| BS-UNet3+ | 0.680 | 0.336 | 0.593 | 0.757 | 0.657 | 0.680 | 45.731 | 14.615 |
| **MS-UNet3+** | **0.909** | **0.752** | **0.842** | **0.918** | **0.901** | **0.909** | **20.475** | **3.339** |
| BS-SwinUNet | 0.394 | 0.367 | 0.045 | 0.547 | 0.435 | 0.415 | 78.911 | 31.017 |
| **MS-SwinUNet** | **0.733** | **0.470** | **0.626** | **0.760** | **0.771** | **0.733** | **44.572** | **10.546** |

**Table 1**. Average segmentation performance of the five clients across different split model architectures for Baseline SplitFed (BS) and MuCALD SplitFed (MS). Best results (MS) are shown in bold.

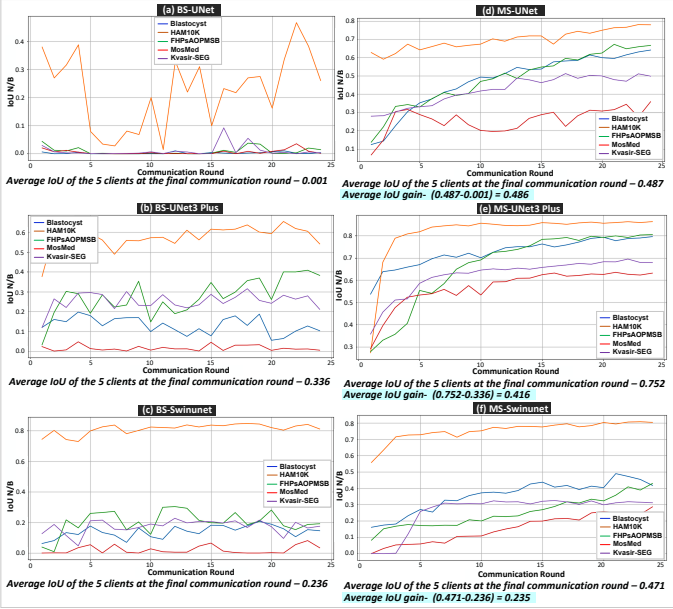



**Fig. 3**. IoU N/B performance across communication rounds for Baseline SplitFed (BS: a,b,c) and MuCALD SplitFed (MS: d,e,f), with average IoU gains reported per model. Average IoU N/B gains: 0.486 for UNet, 0.416 for UNet3 Plus, 0.235 for SwinUNet.

representations. CRDM and DACA provide complementary benefits for stability and cross-client alignment. Although disabling causal modeling increases performance in some cases, it likely reflects reliance on spurious correlations rather than robust generalization. Overall, the full model achieves the best trade-off between accuracy, stability, and privacy. Per-client ablations are available in our code repository.

## 4. CONCLUSION & FUTURE WORKS

This work introduced MuCALD-SplitFed, a causal–latent diffusion framework addressing stability, privacy, and cross-task generalization in multi-task SplitFed. By combining causal representation learning, diffusion-based obfuscation, and domain-adversarial alignment, MuCALD-SplitFed consistently improves segmentation performance across five medical imaging datasets, outperforming Baseline SplitFed and

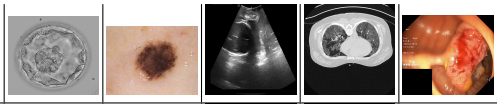
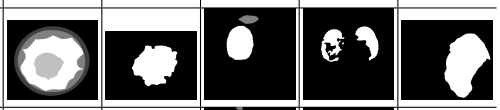
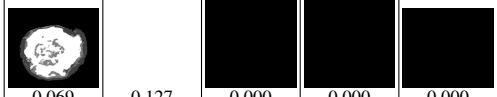
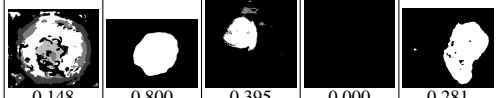
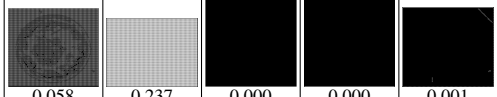
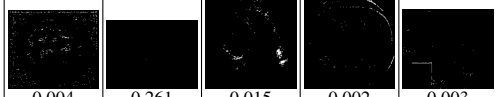
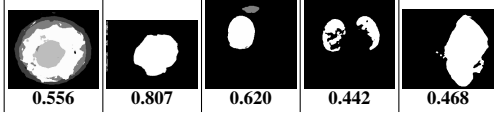
| Method | C#1 | C#2 | C#3 | C#4 | C#5 |
|---|---|---|---|---|---|
| **Sample** |  | | | | |
| **GT** |  | | | | |
| **BS** |  0.044 | 0.161 | 0.007 | 0.011 | 0.002 |
| **FedPer** |  0.069 | 0.127 | 0.000 | 0.000 | 0.000 |
| **FedRep** |  0.148 | 0.800 | 0.395 | 0.000 | 0.281 |
| **FedBN** |  0.179 | 0.121 | 0.086 | 0.000 | 0.000 |
| **FedProx** |  0.162 | 0.433 | 0.068 | 0.130 | 0.131 |
| **SCAFFOLD** |  0.058 | 0.237 | 0.000 | 0.000 | 0.001 |
| **MOCHA** |  0.069 | 0.008 | 0.393 | 0.142 | 0.130 |
| **FedEM** |  0.004 | 0.261 | 0.015 | 0.002 | 0.003 |
| **MS** |  **0.556** | **0.807** | **0.620** | **0.442** | **0.468** |

**Table 2**. Qualitative comparison of MuCALD SplitFed with baseline methods across clients, reporting average Mean IoU N/B per dataset, with best results in bold.

SoTA personalized/ multi-task FL methods. Reconstruction metrics show feature maps are far more obfuscated, making reconstruction and membership-inference attacks effectively infeasible. This work presents an initial step toward comprehensive privacy-preserving multi-task SplitFed learning. Future work will evaluate stronger adversarial models, assess scalability to larger and more diverse client populations, and extend to broader causal settings, additional modalities, and adaptive privacy under increased heterogeneity.

| Method | Client# | Segmentation metrics | | | | | | | | Reconstruction-quality metrics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Dice | IoU W/B | IoU N/B | Precision | Recall | F1 Score | HD95 | ASSD | MSE(S1) | PSNR(S1) | SSIM(S1) | MSE(S2) | PSNR(S2) | SSIM(S2) |
| **BS** | **C#1** | 0.215 | 0.154 | 0.044 | 0.264 | 0.257 | 0.215 | 69.301 | 24.389 | 0.323 | 5.092 | 0.065 | 0.255 | 0.131 | 0.125 |
| | **C#2** | 0.469 | 0.469 | 0.161 | 0.876 | 0.580 | 0.575 | 71.004 | 22.847 | 0.468 | 3.411 | 0.005 | 0.468 | 3.416 | 0.005 |
| | **C#3** | 0.316 | 0.289 | 0.007 | 0.634 | 0.338 | 0.316 | 110.546 | 41.626 | 0.037 | 14.727 | 0.389 | 0.037 | 14.782 | 0.394 |
| | **C#4** | 0.503 | 0.488 | 0.011 | 0.502 | 0.503 | 0.503 | 75.422 | 33.197 | 0.155 | 9.271 | 0.145 | 0.154 | 9.286 | 0.145 |
| | **C#5** | 0.466 | 0.434 | 0.002 | 0.459 | 0.498 | 0.466 | 68.281 | 33.029 | 0.218 | 6.716 | 0.115 | 0.175 | 7.712 | 0.213 |
| | **Avg** | 0.394 | 0.367 | 0.045 | 0.547 | 0.435 | 0.415 | 78.911 | 31.018 | 0.240 | 7.843 | 0.144 | 0.218 | 7.059 | 0.177 |
| **FedPer** | **C#1** | 0.086 | 0.055 | 0.069 | 0.255 | 0.200 | 0.086 | 212.194 | 73.481 | 0.143 | 9.035 | 0.420 | 0.116 | 10.055 | 0.401 |
| | **C#2** | 0.203 | 0.254 | 0.127 | 0.127 | 0.500 | 0.203 | 196.640 | 61.106 | 0.173 | 8.114 | 0.475 | 0.106 | 10.496 | 0.491 |
| | **C#3** | 0.307 | 0.284 | 0.000 | 0.284 | 0.333 | 0.307 | 125.446 | 46.152 | 0.036 | 14.826 | 0.389 | 0.036 | 14.826 | 0.389 |
| | **C#4** | 0.496 | 0.492 | 0.000 | 0.492 | 0.500 | 0.496 | 95.260 | 40.435 | 0.159 | 9.234 | 0.142 | 0.159 | 9.234 | 0.142 |
| | **C#5** | 0.465 | 0.435 | 0.000 | 0.435 | 0.500 | 0.465 | 113.236 | 45.931 | 0.218 | 6.739 | 0.115 | 0.218 | 6.739 | 0.115 |
| | **Avg** | 0.311 | 0.304 | 0.065 | 0.039 | 0.407 | 0.311 | 148.555 | 53.421 | 0.145 | 9.610 | 0.308 | 0.127 | 10.270 | 0.307 |
| **FedRep** | **C#1** | 0.347 | 0.244 | 0.148 | 0.483 | 0.366 | 0.347 | 46.694 | 11.824 | 0.338 | 4.985 | 0.063 | 0.245 | 6.386 | 0.168 |
| | **C#2** | 0.925 | 0.863 | 0.800 | 0.922 | 0.928 | 0.925 | 14.498 | 2.438 | 0.474 | 3.316 | 0.053 | 0.467 | 3.396 | 0.056 |
| | **C#3** | 0.672 | 0.571 | 0.395 | 0.754 | 0.622 | 0.672 | 38.063 | 7.200 | 0.103 | 10.138 | 0.389 | 0.096 | 10.472 | 0.392 |
| | **C#4** | 0.496 | 0.492 | 0.000 | 0.507 | 0.500 | 0.496 | 88.869 | 39.609 | 0.159 | 9.234 | 0.142 | 0.157 | 9.257 | 0.177 |
| | **C#5** | 0.689 | 0.583 | 0.281 | 0.773 | 0.653 | 0.689 | 41.971 | 13.546 | 0.207 | 6.977 | 0.133 | 0.179 | 7.618 | 0.141 |
| | **Avg** | 0.626 | 0.551 | 0.325 | 0.688 | 0.614 | 0.626 | 46.019 | 14.923 | 0.256 | 6.744 | 0.156 | 0.228 | 7.226 | 0.187 |
| **FedBN** | **C#1** | 0.374 | 0.277 | 0.179 | 0.377 | 0.401 | 0.374 | 45.935 | 16.876 | 0.290 | 5.401 | 0.096 | 0.215 | 6.758 | 0.110 |
| | **C#2** | 0.542 | 0.445 | 0.121 | 0.876 | 0.560 | 0.542 | 82.266 | 27.963 | 0.471 | 3.395 | 0.003 | 0.469 | 3.412 | 0.004 |
| | **C#3** | 0.400 | 0.333 | 0.086 | 0.417 | 0.393 | 0.400 | 92.351 | 30.747 | 0.062 | 12.688 | 0.397 | 0.058 | 12.975 | 0.398 |
| | **C#4** | 0.496 | 0.492 | 0.000 | 0.492 | 0.500 | 0.496 | 95.260 | 40.435 | 0.159 | 9.234 | 0.142 | 0.158 | 9.241 | 0.144 |
| | **C#5** | 0.466 | 0.435 | 0.000 | 0.902 | 0.500 | 0.466 | 111.620 | 45.246 | 0.218 | 6.739 | 0.115 | 0.143 | 8.578 | 0.214 |
| | **Avg** | 0.456 | 0.396 | 0.077 | 0.613 | 0.471 | 0.456 | 85.486 | 32.253 | 0.240 | 8.491 | 0.151 | 0.209 | 8.592 | 0.174 |
| **FedProx** | **C#1** | 0.317 | 0.211 | 0.162 | 0.327 | 0.323 | 0.317 | 53.136 | 11.912 | 0.283 | 5.704 | 0.071 | 0.217 | 6.794 | 0.034 |
| | **C#2** | 0.690 | 0.533 | 0.433 | 0.699 | 0.762 | 0.690 | 45.377 | 9.486 | 0.391 | 4.324 | 0.158 | 0.368 | 4.648 | 0.143 |
| | **C#3** | 0.367 | 0.297 | 0.068 | 0.383 | 0.377 | 0.367 | 69.066 | 18.966 | 0.137 | 8.827 | 0.400 | 0.113 | 9.665 | 0.404 |
| | **C#4** | 0.609 | 0.553 | 0.130 | 0.608 | 0.610 | 0.609 | 38.697 | 12.701 | 0.164 | 9.125 | 0.141 | 0.159 | 9.296 | 0.148 |
| | **C#5** | 0.562 | 0.469 | 0.131 | 0.566 | 0.559 | 0.562 | 60.848 | 16.104 | 0.189 | 7.355 | 0.133 | 0.141 | 8.643 | 0.170 |
| | **Avg** | 0.509 | 0.412 | 0.185 | 0.517 | 0.526 | 0.509 | 53.425 | 13.834 | 0.232 | 7.009 | 0.180 | 0.199 | 7.405 | 0.180 |
| **SCAFFOLD** | **C#1** | 0.168 | 0.100 | 0.058 | 0.160 | 0.188 | 0.168 | 104.390 | 41.987 | 0.330 | 4.893 | 0.009 | 0.044 | 14.302 | 0.096 |
| | **C#2** | 0.388 | 0.240 | 0.237 | 0.507 | 0.507 | 0.388 | 9.823 | 52.924 | 0.249 | 6.074 | 0.003 | 0.200 | 7.020 | 0.003 |
| | **C#3** | 0.304 | 0.280 | 0.000 | 0.284 | 0.328 | 0.304 | 105.006 | 53.213 | 0.050 | 13.265 | 0.377 | 0.047 | 13.494 | 0.353 |
| | **C#4** | 0.495 | 0.490 | 0.000 | 0.492 | 0.498 | 0.495 | 77.754 | 53.558 | 0.163 | 9.034 | 0.139 | 0.162 | 9.059 | 0.141 |
| | **C#5** | 0.462 | 0.429 | 0.001 | 0.440 | 0.493 | 0.462 | 72.343 | 34.309 | 0.226 | 6.567 | 0.100 | 0.126 | 9.096 | 0.023 |
| | **Avg** | 0.363 | 0.308 | 0.059 | 0.376 | 0.403 | 0.363 | 73.863 | 47.198 | 0.203 | 7.754 | 0.126 | 0.115 | 10.594 | 0.122 |
| **MOCHA** | **C#1** | 0.217 | 0.164 | 0.037 | 0.350 | 0.249 | 0.217 | 66.205 | 26.965 | 0.334 | 4.995 | 0.069 | 0.171 | 7.936 | 0.230 |
| | **C#2** | 0.620 | 0.503 | 0.218 | 0.875 | 0.608 | 0.620 | 61.653 | 18.819 | 0.473 | 3.364 | 0.008 | 0.472 | 3.376 | 0.008 |
| | **C#3** | 0.420 | 0.352 | 0.102 | 0.576 | 0.405 | 0.420 | 85.198 | 26.946 | 0.060 | 12.985 | 0.393 | 0.059 | 13.079 | 0.366 |
| | **C#4** | 0.496 | 0.492 | 0.000 | 0.491 | 0.500 | 0.496 | 95.260 | 40.435 | 0.159 | 9.234 | 0.142 | 0.159 | 9.235 | 0.142 |
| | **C#5** | 0.602 | 0.503 | 0.172 | 0.618 | 0.592 | 0.602 | 43.753 | 12.004 | 0.219 | 6.708 | 0.130 | 0.154 | 8.224 | 0.141 |
| | **Avg** | 0.471 | 0.403 | 0.106 | 0.582 | 0.471 | 0.471 | 70.414 | 25.034 | 0.248 | 7.257 | 0.148 | 0.159 | 8.770 | 0.178 |
| **FedEM** | **C#1** | 0.178 | 0.074 | 0.004 | 0.160 | 0.178 | 0.152 | 107.140 | 41.681 | 0.434 | 3.821 | 0.001 | 0.260 | 6.462 | 0.096 |
| | **C#2** | 0.473 | 0.517 | 0.261 | 0.507 | 0.518 | 0.368 | 9.682 | 53.124 | 0.471 | 3.375 | 0.009 | 0.329 | 5.312 | 0.057 |
| | **C#3** | 0.150 | 0.291 | 0.015 | 0.284 | 0.314 | 0.314 | 103.006 | 54.213 | 0.032 | 15.290 | 0.393 | 0.014 | 18.840 | 0.303 |
| | **C#4** | 0.495 | 0.486 | 0.002 | 0.492 | 0.499 | 0.494 | 76.754 | 55.558 | 0.169 | 8.679 | 0.127 | 0.112 | 10.174 | 0.063 |
| | **C#5** | 0.513 | 0.435 | 0.003 | 0.440 | 0.495 | 0.482 | 70.344 | 31.309 | 0.219 | 6.708 | 0.112 | 0.163 | 7.993 | 0.098 |
| | **Avg** | 0.362 | 0.361 | 0.057 | 0.377 | 0.401 | 0.362 | 73.385 | 47.177 | 0.264 | 7.774 | 0.187 | 0.156 | 9.297 | 0.124 |
| **MS** | **C#1** | **0.728** | **0.574** | **0.556** | **0.739** | **0.721** | **0.728** | **45.730** | **6.717** | **0.430** | **35.264** | **0.790** | **0.110** | **35.544** | **0.748** |
| | **C#2** | **0.929** | **0.869** | **0.807** | **0.930** | **0.927** | **0.929** | **26.008** | **3.835** | **0.238** | **32.753** | **0.467** | **0.071** | **40.285** | **0.764** |
| | **C#3** | **0.831** | **0.726** | **0.620** | **0.867** | **0.816** | **0.831** | **47.909** | **8.815** | **0.314** | **38.560** | **0.770** | **0.082** | **40.298** | **0.816** |
| | **C#4** | **0.802** | **0.713** | **0.442** | **0.755** | **0.874** | **0.802** | **28.834** | **5.003** | **0.422** | **36.594** | **0.663** | **0.119** | **35.709** | **0.742** |
| | **C#5** | **0.790** | **0.680** | **0.468** | **0.780** | **0.802** | **0.790** | **37.762** | **6.646** | **0.266** | **34.302** | **0.561** | **0.092** | **38.277** | **0.756** |
| | **Avg** | **0.816** | **0.712** | **0.579** | **0.814** | **0.828** | **0.816** | **37.249** | **6.203** | **0.334** | **35.695** | **0.650** | **0.095** | **38.823** | **0.765** |

**Table 3**. Segmentation and reconstruction-quality metrics over the 5 clients for all methods. BS: Baseline SplitFed, MS: MuCALD SplitFed, S1: Split-1, S2: Split-2, Best values (MS) are in bold.

| Ablation study | Segmentation metrics | | | | | | | | Reconstruction-quality metrics | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Dice | IoU W/B | IoU N/B | Precision | Recall | F1 Score | HD95 | ASSD | MSE(S1) | PSNR(S1) | SSIM(S1) | MSE(S2) | PSNR(S2) | SSIM(S2) |
| **CRDM only** | 0.814 | 0.720 | 0.579 | 0.858 | 0.801 | 0.814 | 27.957 | 5.551 | 1.175 | 53.351 | 0.922 | 0.116 | 36.608 | 0.758 |
| **DACA only** | 0.832 | 0.734 | 0.607 | 0.850 | 0.837 | 0.832 | 30.489 | 5.272 | — | — | — | — | — | — |
| **Causality disabled** | 0.843 | 0.747 | 0.627 | 0.857 | 0.844 | 0.843 | 31.544 | 5.251 | 1.151 | 53.093 | 0.940 | 0.132 | 36.120 | 0.771 |
| **Diffusion disabled** | 0.800 | 0.689 | 0.552 | 0.811 | 0.817 | 0.800 | 44.269 | 7.264 | — | — | — | — | — | — |
| **Forward noise disabled** | 0.785 | 0.683 | 0.533 | 0.808 | 0.792 | 0.785 | 46.239 | 8.790 | — | — | — | — | — | — |

**Table 4**. Average segmentation performance for ablation experiments with MS-UNet. S1: Split-1, S2: Split-2. Reconstruction metrics are not applicable in some cases due to architectural modifications in the ablated models.

## 5. REFERENCES


[1] Brendan McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] Otkrist Gupta et al., "Distributed learning of deep neural network over multiple agents," *J. Netw. Comput.*, vol. 116, pp. 1–8, Aug. 2018.

[3] Chandra Thapa et al., "SplitFed: When Federated Learning Meets Split Learning," in *Proc. AAAI*, 2022, vol. 36, pp. 8485–8493.

[4] Virginia Smith et al., "Federated multi-task learning," *Proc. NeurIPS*, vol. 30, 2017.

[5] Chaouki Ben Issaid et al., "Tackling feature and sample heterogeneity in decentralized multi-task learning: A sheaf-theoretic approach," *arXiv preprint arXiv:2502.01145*, 2025.

[6] Sunny Gupta et al., "Fedalign: Federated domain generalization with cross-client feature alignment," *arXiv preprint arXiv:2501.15486*, 2025.

[7] Yi-Qiang Chen et al., "Fedbone: Towards large-scale federated multi-task learning," *JCST*, vol. 39, no. 5, pp. 1040–1057, 2024.

[8] Jiayi Chen et al., "Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks," in *Proc. ACM SIGKDD*, 2022, pp. 87–96.

[9] Yipeng Dong et al., "Multi-task federated split learning across multi-modal data with privacy preservation," *Sensors*, vol. 25, no. 1, pp. 233, 2025.

[10] Judea Pearl et al., *The book of why: the new science of cause and effect*, Basic books, 2018.

[11] Daniel C Castro et al., "Causality matters in medical imaging," *Nat. Commun.*, vol. 11, no. 1, pp. 3673, 2020.

[12] Zhang Chen et al., "C-cam: Causal cam for weakly supervised semantic segmentation on medical image," in *Proc. CVPR*, 2022, pp. 11676–11685.

[13] Juzheng Miao et al., "Caussl: Causality-inspired semi-supervised learning for medical image segmentation," in *Proc. CVPR*, 2023, pp. 21426–21437.

[14] Vibujithan Vigneshwaran et al., "Macaw: A causal generative model for medical imaging," *arXiv preprint arXiv:2412.02900*, 2024.

[15] Aneesh Komanduri et al., "Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models," in *ECAI 2024*, pp. 2516–2523. IOS Press, 2024.

[16] Pedro Sanchez et al., "Diffusion causal models for counterfactual estimation," in *Conference on Causal Learning and Reasoning*. PMLR, 2022, pp. 647–668.

[17] Mingkun Zhang et al., "Causaldiff: Causality-inspired disentanglement via diffusion model for adversarial defense," in *Proc. NeurIPS*, 2025.

[18] Chamani Shiranthika et al., "Decentralized learning in healthcare: A review of emerging techniques," *IEEE Access*, vol. 11, pp. 54188–54209, 2023.

[19] Chamani Shiranthika et al., "Adaptive asynchronous split federated learning for medical image segmentation," *IEEE Access*, vol. 12, pp. 182496–182515, 2024.

[20] Xun Zheng et al., "Learning sparse nonparametric dags," in *Proc. ICML*. PMLR, 2020, pp. 3414–3425.

[21] Lisette Lockhart et al., "Multi-Label Classification for Automatic Human Blastocyst Grading with Severely Imbalanced Data," in *Proc. MMSP*, Kuala Lumpur, Malaysia, Sept. 2019, pp. 1–6.

[22] Yaosheng Lu et al., "The jnu-ifm dataset for segmenting pubic symphysis-fetal head," *Data Br*, vol. 41, pp. 107904, 2022.

[23] Philipp Tschandl et al., "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

[24] Sergey P Morozov et al., "Mosmeddata: data set of 1110 chest ct scans performed during the covid-19 epidemic," *Digit. Diagn.*, vol. 1, no. 1, pp. 49–59, 2020.

[25] Debesh Jha et al., "Kvasir-seg: A segmented polyp dataset," in *Proc. MMM*. Springer, 2020, pp. 451–462.

[26] Yaroslav Ganin et al., "Unsupervised domain adaptation by backpropagation," in *ICML*. PMLR, 2015, pp. 1180–1189.

[27] Carole H. Sudre et al., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. DLMIA*, M. Jorge Cardoso et al., Eds., Cham, 2017, pp. 240–248, Springer International Publishing.

[28] Michael A. A. Cox et al., *Multidimensional Scaling*, pp. 315–347, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[29] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Springer, 2015, pp. 234–241.

[30] Huimin Huang et al., "Unet 3+: A full-scale connected unet for medical image segmentation," in *Proc. IEEE ICASSP*. Ieee, 2020, pp. 1055–1059.

[31] Hu Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. ECCV*. Springer, 2022, pp. 205–218.

[32] Manoj Ghuhan Arivazhagan et al., "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[33] Liam Collins et al., "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*. PMLR, 2021, pp. 2089–2099.

[34] Xiaoxiao Li et al., "Fedbn: Federated learning on non-iid features via local batch normalization," in *Proc. ICML*, 2021.

[35] Tian Li et al., "Federated optimization in heterogeneous networks," *MLSys*, vol. 2, pp. 429–450, 2020.

[36] Sai Praneeth Karimireddy et al., "Scaffold: Stochastic controlled averaging for federated learning," in *ICML*. PMLR, 2020, pp. 5132–5143.

[37] Othmane Marfoq et al., "Federated multi-task learning under a mixture of distributions," *NeurIPS*, vol. 34, pp. 15434–15447, 2021.