



Supplementary Material

Training details and hyperparameters

The training configuration was established using a hand-tuned, one-factor-at-a-time ablation on the pooled three-site data, initialized from the publicly released GMIC weights (ensemble member `sample_model_5`). The ablation evaluated class-imbalance handling, differential learning rates, and the backbone-freezing schedule. The resulting configuration was then held fixed across the local baselines and federated experiments.

For each training example, the loss comprised the sum of binary cross-entropy terms for the global, local, and fusion outputs, plus β times the L_1 norm of the malignant saliency map. Class imbalance (view-level positive-to-negative ratio 7.2:1) was addressed using a class-balanced sampler with a positive-class weight of 1.0. Models were optimized with Adam using learning rates of 3×10^{-5} for the classification heads and 3×10^{-7} for the backbone, weight decay 1×10^{-6} , and an initial backbone freeze of 2 epochs. Local models were trained for up to 40 epochs with early stopping on breast-level validation AUC (patience 6) and batch size 32. Federated methods were trained for 60 rounds with one local epoch per round. FedProx used $\mu = 0.1$.

Ditto regularization sweeps (Table S1)

Table S1. Ditto validation-AUC sweeps (12 communication rounds; model selection only): worst-site (always Site B, RSNA-US) and mean per-site AUC. All variation is within the validation CI (≤ 0.007 AUC); we adopt scalar $\lambda = 0.05$.

(a) Scalar λ sweep.

λ	worst	mean
0.01	0.8556	0.8948
0.05	0.8562	0.8945
0.1	0.8560	0.8944
0.5	0.8549	0.8923
1	0.8540	0.8909
2	0.8521	0.8901

(b) Module-wise sweep.

shared group	worst	mean
anchor (all 0.01)	0.8551	0.8945
global 0.5	0.8556	0.8942
global 2	0.8531	0.8923
local 0.5	0.8547	0.8937
local 2	0.8540	0.8930
fusion 0.5	0.8536	0.8926
fusion 2	0.8516	0.8911

Per-site test AUC at each site’s own best-validation round (Table S2)

Table S2. Per-site test AUC for the federated methods at each site’s *own* best-validation round (round in parentheses); the Centralized and Local-only baselines (no round) are in Table 2 of the main paper. Complements the main-paper table, which fixes one common pooled-validation round.

Strategy	Site A (HIPIMR)	Site B (RSNA-US)	Site C (RSNA-AU)
FedAvg	0.898 (r58)	0.842 (r0)	0.856 (r2)
FedProx	0.905 (r55)	0.842 (r0)	0.858 (r2)
FedBN	0.899 (r37)	0.842 (r0)	0.831 (r1)
Ditto	0.899 (r54)	0.862 (r42)	0.903 (r46)
Module-wise Ditto	0.900 (r8)	0.863 (r46)	0.887 (r35)

Per-group performance across methods on Site A (Table S3)

Table S3. Site A (HIPIMR) per-group breast-level AUC by strategy, each deployed at its pooled-validation round (Pretrained is the un-fine-tuned baseline). Complements Table 3 of the main paper (Pretrained vs. Ditto). Every federated strategy lifted NHPI well above the pretrained baseline, with Ditto marginally highest; on this best-served site the methods were otherwise comparable. Group sizes n (positive breasts) are as in Table 3; [†]small- n groups are exploratory. Per-group operating-point metrics for every method are in the released analysis.

Strategy	Overall	NHPI [†]	Asian	White [†]
Pretrained	0.870	0.839	0.882	0.825
FedAvg	0.898	0.896	0.905	0.855
FedProx	0.905	0.903	0.911	0.865
FedBN	0.899	0.888	0.911	0.838
Ditto	0.899	0.907	0.901	0.877
Module-wise Ditto	0.900	0.898	0.904	0.867