arXiv:2608.09240v1 [cs.LG] 10 Aug 2026

Multimodal Federated Learning under Dual-Axis Modality Missingness

Adiba Orzikulova[*1] **Jaehyun Kwak**[*1]
Jaemin Shin[1] **Yunqi Guo**[2] **Xiaomin Ouyang**[3] **Guoliang Xing**[2]
Steven Euijong Whang[1] **Sung-Ju Lee**[1]
[1]KAIST, [2]CUHK, [3]HKUST
{adiorz,jaehyun98,jaemin.shin,swhang,profsj}@kaist.ac.kr
yunqiguo@cuhk.edu.hk, glxing@ie.cuhk.edu.hk
xmouyang@cse.ust.hk

Abstract

Multimodal federated learning (FL) supports collaborative modeling in privacy-sensitive health-sensing and medical settings, but realistic deployments often exhibit *dual-axis modality missingness*: clients have different modality sets, and individual samples may contain only subsets of the modalities available locally. Existing methods typically address these two axes separately. We propose `Flux`, a multimodal federated learning framework built around two complementary components. First, *modality-aware confidence tempering* learns sample-specific confidence for each modality through mask-aware unimodal supervision and fuses the confidence estimates from observed modalities into a sample-adaptive temperature that adjusts predictive sharpness according to evidence quality and completeness. Second, *gradient-decoupled private adaptation* applies this temperature only to a client-private prediction pathway, while training the shared federated model with a standard, untempered objective. This enables sample-specific, client-local confidence adaptation without allowing confidence-dependent gradients to perturb shared representation learning. Across four multimodal datasets, `Flux` achieves the highest average macro-F1 on every dataset, outperforming the strongest dataset-specific baseline by 0.8$\sim$2.2 points and by 1.6 points on average. Additional analyses demonstrate favorable calibration, temperature sensitivity to both modality missingness and input corruption, and more stable shared optimization under private-only tempering. Our code is available at https://github.com/AdibaOrz/Flux.

1 Introduction

Multimodal deep learning is increasingly deployed in privacy- and regulation-constrained settings, including medical and health-sensing applications [27, 8, 26, 34, 21]. In these settings, data are often distributed across clients or silos such as hospitals, personal devices, or deployment sites, where centralizing raw data may be impractical or prohibited. Federated learning (FL) therefore provides a practical framework for collaborative model training without exchanging raw data [17, 11].

A central challenge in realistic multimodal FL is modality missingness along two distinct but coupled axes. At the client level, clients may have access to different sets of modalities, a setting we refer to as *inter-client modality heterogeneity*. For example, hospitals and devices may collect different subsets of imaging, genomic, clinical, or sensor modalities due to differences in acquisition protocols, infrastructure, or device capabilities. At the sample level, a particular instance may contain only a subset of the modalities available at its client, yielding *intra-client modality missingness*. For example, a patient record may lack a modality because an examination was not performed, a test was cost-prohibitive, or data acquisition failed. These two forms of missingness often co-occur

[*]Equal contribution

Preprint.

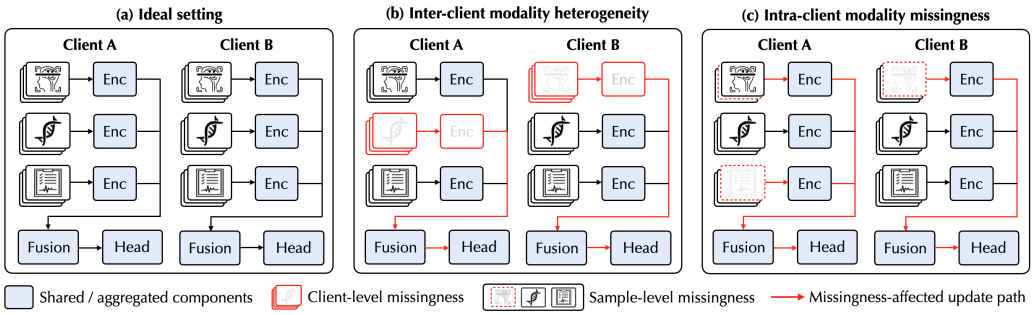

Figure 1: *Dual-axis modality missingness* in multimodal FL: (a) the complete-modality setting, (b) *inter-client modality heterogeneity*, and (c) *intra-client modality missingness*. Red arrows mark missingness-affected update paths.

in realistic deployments [31, 30], creating a *dual-axis modality missingness* regime. This regime violates the common simplifying assumption in multimodal FL that all clients share the same modality schema and that every sample contains complete observations for all modalities [33]. Figure 1 contrasts this idealized setting (a) with inter-client modality heterogeneity (b) and intra-client modality missingness (c).

Existing methods generally address these two axes separately. Approaches to inter-client modality heterogeneity coordinate learning across clients with different modality sets through modality-wise training or quality-aware aggregation [22, 20], but assume complete observations within each client. Conversely, methods for intra-client modality missingness improve robustness to partially observed samples, often through masking [3, 6], while assuming a common modality schema across clients. These assumptions overlook their interaction during federated optimization: sample-level missingness determines which representations contribute to each local update, whereas client-level availability determines which modality branches and combinations each client can train. The server therefore aggregates shared components optimized under systematically different evidence distributions and modality coverage.

To address this coupled challenge, we propose `Flux`, which separates sample-level evidence adaptation from shared federated optimization. First, *modality-aware confidence tempering* learns task-relevant confidence for each observed modality through mask-aware unimodal supervision and combines these estimates into a sample-adaptive temperature. The temperature adjusts predictive sharpness when a sample contains fewer, noisier, or less informative modalities, thereby addressing intra-client variation in evidence completeness and quality.

Applying this confidence adjustment directly to the shared pathway, however, would introduce additional confidence-dependent variation into shared gradients. Because clients differ in both modality availability and sample-level missingness patterns, temperature-dependent changes in gradient magnitude and direction could amplify heterogeneity among the updates aggregated by the server. *Gradient-decoupled private adaptation* prevents this coupling by applying the temperature only to a client-private prediction head operating on a detached fused representation, while training the shared encoders, fusion module, and global predictor with an untempered objective. Thus, confidence tempering guides evidence-aware private adaptation, whereas gradient decoupling prevents these sample-specific adjustments from distorting collaboratively learned representations.

We evaluate `Flux` on four multimodal datasets spanning health-sensing and medical domains: PAMAP2 [24], RealWorldHAR [28], SleepEDF [7, 12], and ADNI [32], a heterogeneous biomedical dataset with naturally missing modalities. `Flux` achieves the highest average macro-F1 on each dataset, outperforming the strongest dataset-specific baseline by 0.8$\sim$2.2 points, with an average gain of 1.6 points across datasets. Beyond predictive performance, our analyses show that `Flux` achieves favorable calibration and learns a confidence-derived temperature that increases with modality incompleteness and input corruption. Private-only tempering also reduces shared-gradient instability and client-to-global update drift.

In summary, this work makes three contributions. First, we formulate *dual-axis modality missingness* in federated multimodal learning as a coupled regime in which client-level modality availability and sample-level modality observation jointly shape local optimization and server-side aggregation. Second, we propose `Flux`, which combines modality-aware confidence tempering with gradient-

decoupled private adaptation, forming a sample-adaptive temperature from per-modality confidence while isolating confidence-tempered updates from shared parameters. Third, we demonstrate consistent performance gains across four health-sensing and medical datasets, including naturally incomplete biomedical data, and provide diagnostic analyses of calibration, sensitivity to evidence quality and shared-optimization stability.

2 Problem Formulation

We consider a federated multimodal learning setting with K clients and a universal modality set $\mathcal{M} = \{1, \ldots, M\}$. Each client $k \in \{1, \ldots, K\}$ holds a private dataset:

$$\mathcal{D}_k = \{(\mathbf{x}_{k,i}, y_{k,i}, \mathbf{a}_{k,i})\}_{i=1}^{n_k}.$$

For a sample i on client k, $\mathbf{x}_{k,i} = (x_{k,i,1}, \ldots, x_{k,i,M})$ is the multimodal input, $y_{k,i}$ is the label, and $\mathbf{a}_{k,i} \in \{0,1\}^M$ is the sample-level observation mask, where $a_{k,i,m} = 1$ indicates that modality m is observed for that sample. We further define a client-level availability mask $\mathbf{c}_k \in \{0,1\}^M$, where $c_{k,m} = 1$ indicates that modality m is available at client k. A sample can contain only modalities available at its client:

$$a_{k,i,m} \le c_{k,m}, \quad \forall k \in \{1, \ldots, K\}, i \in \{1, \ldots, n_k\}, m \in \mathcal{M}.$$

We assume that every sample contains at least one observed modality, i.e., $\sum_{m=1}^{M} a_{k,i,m} \ge 1$. Variation in $\{\mathbf{c}_k\}_{k=1}^{K}$ characterizes *inter-client modality heterogeneity*. In particular, clients k and k' have different modality profiles when $\mathbf{c}_k \neq \mathbf{c}_{k'}$. By contrast, $c_{k,m} = 1$ and $a_{k,i,m} = 0$ indicate *intra-client modality missingness*: a modality m is available at client k but absent from sample i.

The goal is to learn a federated predictor that performs well when these two forms of missingness occur jointly. Let θ denote the shared parameters communicated to and aggregated by the server, and let ϕ_k denote the parameters maintained privately by client k. For methods without client-specific components, ϕ_k is omitted. The local empirical objective for client k is

$$\mathcal{L}_k(\theta, \phi_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(f(\mathbf{x}_{k,i}, \mathbf{a}_{k,i}; \theta, \phi_k), y_{k,i}),$$

where $f(\cdot)$ is the multimodal predictor and $\ell(\cdot, \cdot)$ is the task loss. The corresponding federated objective is

$$\min_{\theta, \{\phi_k\}_{k=1}^{K}} \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta, \phi_k), \qquad n = \sum_{k=1}^{K} n_k.$$

At communication round t, the server distributes the current shared parameters θ^t to a selected client set $\mathcal{S}^t$. Each selected client performs local optimization and returns its updated shared parameters θ_k^{t+1} while retaining its private parameters locally. The server then computes

$$\theta^{t+1} = \mathrm{Agg}\big(\{\theta_k^{t+1} : k \in \mathcal{S}^t\}\big),$$

where $\mathrm{Agg}(\cdot)$ denotes an FL aggregation rule, such as sample-size-weighted averaging [17].

Under dual-axis modality missingness, the two masks affect optimization at different granularities. The sample-level mask $\mathbf{a}_{k,i}$ determines which modality evidence contributes to each sample loss and the resulting gradient, whereas the client-level mask $\mathbf{c}_k$ constrains which modality branches and cross-modal combinations client k can optimize. Consequently, the server aggregates shared components trained under client-dependent modality coverage and evidence distributions. This creates two coupled requirements: adapting each prediction to the evidence observed for that sample, and preventing client-specific modality availability and missingness from destabilizing shared federated optimization. `Flux` addresses them through modality-aware confidence tempering and gradient-decoupled private adaptation, respectively.

3 `Flux` Design

Figure 2 summarizes `Flux`. For each sample, shared modality encoders transform the observed modalities into unimodal representations. *Modality-aware confidence tempering* estimates confidence

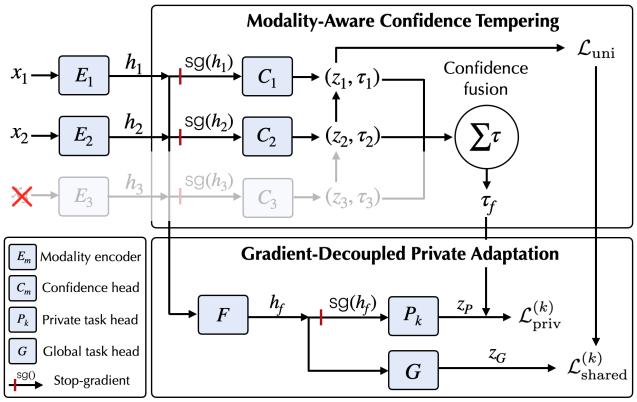

Figure 2: Overview of `Flux`. Faded paths denote missing modalities. Observed-modality temperatures are fused into τ_f, which is applied to the private pathway operating on a detached fused representation.

for each observed modality, and combines these estimates into a sample-adaptive temperature τ_f that reflects the completeness and reliability of the observed evidence. *Gradient-decoupled private adaptation* then separates collaborative representation learning from confidence-dependent private adaptation: the shared multimodal pathway is trained with a standard, untempered task objective, whereas τ_f is applied only to the logits of a client-private task head operating on a detached fused representation. The following subsections describe these components, followed by the federated training and inference procedures. The training algorithm of `Flux` is in Appendix A.

3.1 Modality-Aware Confidence Tempering

Samples can vary not only in the number of observed modalities, but also in the reliability of their observed evidence. `Flux` captures this variation by estimating an inverse-confidence temperature for each observed modality and combining the estimates into a sample-adaptive temperature. Unlike post-hoc sample-dependent temperature scaling [10], which predicts a single test-time temperature for the final logits of a fixed classifier, `Flux` learns modality-level temperature during training and uses their fused value only to guide client-private adaptation.

We describe the method for a single training sample with label y and observation mask $\mathbf{a} \in \{0,1\}^M$, omitting client and sample indices for clarity. For each observed modality m, i.e., $a_m = 1$, an encoder E_m produces

$$h_m = E_m(x_m).$$

Per-modality confidence. A modality-specific auxiliary confidence head C_m receives the detached unimodal representation and outputs unimodal logits z_m and a scalar log-scale parameter ν_m:

$$(z_m, \nu_m) = C_m(\mathrm{sg}(h_m)), \tag{1}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Inspired by input-dependent uncertainty parametrization [13], we define the positive unimodal temperature

$$\tau_m = \exp(\nu_m/2).$$

We use τ_m as a deterministic inverse-confidence proxy: scaling the logits as z_m/τ_m sharpens the predictive distribution for lower τ_m and flattens it for higher τ_m.

Mask-aware unimodal supervision. Let $N_{\mathbf{a}} = \sum_{m=1}^{M} a_m$ denote the number of observed modalities. We train confidence heads of the observed modalities using a temperature-scaled unimodal classification loss:

$$\mathcal{L}_{\text{uni}} = \frac{1}{N_{\mathbf{a}}} \sum_{m=1}^{M} a_m \,\mathrm{CE}\left(\frac{z_m}{\tau_m}, y\right). \tag{2}$$

Since h_m is detached, $\mathcal{L}_{\text{uni}}$ updates only $\{C_m\}$. This confines the auxiliary unimodal objective to the confidence heads, leaving the shared encoders optimized for multimodal fusion.

Fused multimodal temperature. Treating τ_m^2 as a modality-level relative uncertainty proxy, we combine the observed estimates using precision-inspired aggregation:

$$\tau_f = \left(\sum_{m=1}^{M} \frac{a_m}{\tau_m^2 + \epsilon} + \epsilon \right)^{-1/2}, \tag{3}$$

where $\epsilon > 0$ ensures numerical stability. Lower-temperature modalities contribute more strongly to the aggregation, while fewer or less reliable observations yield a larger fused temperature. Thus, τ_f reflects both the completeness and estimated reliability of the observed evidence.

Applying τ_f directly to the shared pathway would introduce additional confidence-dependent variation into shared gradients tied to client-specific observation patterns. This motivates the gradient separation introduced next.

3.2 Gradient-Decoupled Private Adaptation

`Flux` separates shared task learning from confidence-tempered client-private adaptation during local optimization. The shared path is trained with untempered task logits, whereas τ_f scales only the logits of a private head operating on a detached fused representation. This confines confidence-tempered gradients to the private head and prevents them from affecting parameters aggregated by the server.

Shared pathway. A mask-aware fusion module F combines only the observed unimodal representations:

$$h_f = F\big(\{(a_m, h_m)\}_{m=1}^{M}\big).$$

The shared task head G then produces

$$z_G = G(h_f).$$

On client k, the objective to update shared parameters is

$$\mathcal{L}_{\text{shared}}^{(k)} = \mathbb{E}_{(\mathbf{x},y,\mathbf{a})\sim\mathcal{D}_k} \left[\text{CE}(z_G, y) + \mathcal{L}_{\text{uni}} \right]. \tag{4}$$

The untempered task loss $\text{CE}(z_G, y)$ updates the shared encoders, fusion module, and task head. Because h_m is detached in $\mathcal{L}_{\text{uni}}$, the auxiliary loss updates only the confidence heads and does not push the shared modality encoders toward unimodal prediction.

Private pathway. Each client k maintains a private task head P_k with parameters ϕ_k. The head operates on the detached fused representation:

$$z_{P,k} = P_k(\text{sg}(h_f)). \tag{5}$$

The private objective scales these logits using the detached fused temperature:

$$\mathcal{L}_{\text{priv}}^{(k)} = \mathbb{E}_{(\mathbf{x},y,\mathbf{a})\sim\mathcal{D}_k} \left[\text{CE}\left(\frac{z_{P,k}}{\text{sg}(\tau_f)}, y \right) \right]. \tag{6}$$

Thus, τ_f modulates the private-head training according to the estimated completeness and reliability of the observed evidence, while $\mathcal{L}_{\text{priv}}^{(k)}$ updates only ϕ_k. Detaching h_f blocks gradients to the modality encoders and fusion module, while detaching τ_f blocks gradients to the confidence heads.

Rationale for private-only tempering. If τ_f were applied to the shared logits, then, treating it as constant, the gradient with respect to z_G would be

$$\frac{\partial}{\partial z_G} \text{CE}\left(\frac{z_G}{\tau_f}, y \right) = \frac{1}{\tau_f} \left(p_\tau - e_y \right),$$

where $p_\tau = \text{softmax}(z_G/\tau_f)$ and e_y is the one-hot label vector. With $p = \text{softmax}(z_G)$, this can be written as:

$$\frac{1}{\tau_f}(p_\tau - e_y) = \frac{1}{\tau_f}(p - e_y) + \frac{1}{\tau_f}(p_\tau - p).$$

Relative to the untempered gradient $p - e_y$, τ_f both rescales the gradient and introduces a temperature-dependent correction. In multiclass settings, this correction is generally not collinear with the original gradient, so tempering can alter both its magnitude and direction. Because τ_f varies with each sample's observed evidence, applying it to the shared pathway would introduce additional sample-dependent variation into the updates aggregated across clients. `Flux` therefore confines τ_f to the private-head loss, preventing temperature-dependent modification of shared task gradients.

3.3 Federated Training and Inference

Local training. At communication round t, each selected client k sets its local shared parameters to $\theta^{(t)}$ and loads its retained private parameters ϕ_k. For each minibatch, the client computes the unimodal representations, modality-level confidence outputs, fused representation h_f, fused temperature τ_f, and shared and private logits. It updates the shared parameters using $\mathcal{L}^{(k)}_{\text{shared}}$ and the private parameters ϕ_k using $\mathcal{L}^{(k)}_{\text{priv}}$. After local optimization, client k returns its updated shared parameters $\theta^{(t+1)}_k$ to the server and retains the updated ϕ_k locally for the next round.

Server aggregation. Let

$$n_{k,m} = \sum_{i=1}^{n_k} a_{k,i,m}$$

denote the number of samples in client k's local training set for which modality m is observed, and define

$$\mathcal{S}^t_m = \{k \in \mathcal{S}^t : n_{k,m} > 0\}$$

as the selected clients that contribute an update for modality m. For each modality-specific parameter block

$$\theta_m = (\theta_{E_m}, \theta_{C_m}),$$

containing the parameters of encoder E_m and confidence head C_m, the server computes:

$$\theta^{(t+1)}_m = \begin{cases} \displaystyle\sum_{k \in \mathcal{S}^t_m} \frac{n_{k,m}}{\sum_{j \in \mathcal{S}^t_m} n_{j,m}} \theta^{(t+1)}_{k,m}, & \mathcal{S}^t_m \neq \emptyset, \\ \theta^{(t)}_m, & \mathcal{S}^t_m = \emptyset. \end{cases}$$

The remaining shared parameter blocks, namely the fusion module and shared task head, are aggregated using client-level sample weights:

$$\theta^{(t+1)}_{F,G} = \sum_{k \in \mathcal{S}^t} \frac{n_k}{\sum_{j \in \mathcal{S}^t} n_j} \theta^{(t+1)}_{k,F,G},$$

where $\theta_{F,G}$ contains the parameters of F and G. The private parameters $\{\phi_k\}$ are never communicated or aggregated.

Inference. At test time, client k combines the shared and private logits to obtain the predictive distribution:

$$\hat{z}_k = z_G + z_{P,k}, \qquad \hat{p}_k = \text{softmax}(\hat{z}_k), \hat{y}_k \quad = \arg\max_c [\hat{p}_k]_c.$$

We do not train the model using a joint loss on $z_G + z_{P,k}$. Such a loss would make the gradient of the shared head depend on the client-private logits, coupling updates to the shared pathway with each client's private model state. The confidence heads and τ_f are used only to guide private-head optimization during training and are not required at inference. Confidence estimation is therefore learned collaboratively, while confidence-tempered adaptation remains client-local. Relative to the shared model, inference requires only the lightweight private task head.

4 Experiments

Datasets and baselines. We evaluate `Flux` across four multimodal datasets spanning health-sensing and medical domains: PAMAP2 [24], RealWorldHAR [28], SleepEDF [7, 12], and data obtained from Alzheimer's Disease NeuroImaging Initiative (ADNI) database [32], which contains naturally incomplete multimodal records. For the three health-sensing datasets, we construct controlled dual-axis missingness regimes. For ADNI, we retain the naturally occurring sample-level missingness and introduce inter-client heterogeneity by removing selected modalities from all samples at each client. We benchmark `Flux` against baselines from four methodological families: the canonical FL method FedAvg [17]; non-IID–robust methods FedProx [16] and MOON [15]; personalization and decoupling approaches FedPer [1] and FedRoD [4]; and missing-modality FL methods PmcmFL [3] and PEPSY [18]. Details of the dual-axis missingness construction are provided in Appendix B, while additional information on the datasets, baselines, training, and evaluation protocols is provided in Appendix D.

Table 1: Macro-F1 under dual-axis missingness on three health-sensing datasets. **H** and **I** denote inter-client heterogeneity levels {Homogeneous, Moderate, High} and intra-client incompleteness levels {Moderate, High}.

PAMAP2								
(H, I)	**FedAvg**	**FedProx**	**MOON**	**FedPer**	**FedRoD**	**PmcmFL**	**PEPSY**	**Flux (Ours)**
(Hom, Mod)	0.722 ± 0.007	0.722 ± 0.003	0.723 ± 0.011	0.694 ± 0.012	0.749 ± 0.009	0.723 ± 0.008	0.722 ± 0.015	**0.761 ± 0.006**
(Hom, High)	0.596 ± 0.025	0.600 ± 0.022	0.610 ± 0.012	0.453 ± 0.011	0.609 ± 0.012	0.605 ± 0.007	0.590 ± 0.013	**0.641 ± 0.010**
(Mod, Mod)	0.634 ± 0.008	0.636 ± 0.004	0.636 ± 0.014	0.644 ± 0.013	0.664 ± 0.012	0.643 ± 0.015	0.641 ± 0.014	**0.683 ± 0.012**
(Mod, High)	0.574 ± 0.008	0.569 ± 0.013	0.561 ± 0.018	0.490 ± 0.006	0.588 ± 0.013	0.578 ± 0.020	0.539 ± 0.009	**0.596 ± 0.024**
(High, Mod)	0.621 ± 0.008	0.620 ± 0.011	0.616 ± 0.007	0.614 ± 0.011	0.636 ± 0.011	0.618 ± 0.006	0.627 ± 0.017	**0.655 ± 0.006**
(High, High)	0.499 ± 0.014	0.497 ± 0.016	0.494 ± 0.014	0.440 ± 0.009	0.500 ± 0.017	0.514 ± 0.019	0.508 ± 0.011	**0.520 ± 0.014**
Average	0.608 ± 0.011	0.607 ± 0.011	0.607 ± 0.013	0.556 ± 0.010	0.624 ± 0.012	0.614 ± 0.012	0.604 ± 0.013	**0.642 ± 0.012**

SleepEDF								
(H, I)	**FedAvg**	**FedProx**	**MOON**	**FedPer**	**FedRoD**	**PmcmFL**	**PEPSY**	**Flux (Ours)**
(Hom, Mod)	0.589 ± 0.010	0.589 ± 0.007	0.594 ± 0.008	0.580 ± 0.011	0.600 ± 0.009	0.601 ± 0.006	0.589 ± 0.006	**0.616 ± 0.008**
(Hom, High)	0.526 ± 0.007	0.534 ± 0.007	0.527 ± 0.007	0.510 ± 0.004	0.523 ± 0.012	0.547 ± 0.005	0.537 ± 0.015	**0.557 ± 0.007**
(Mod, Mod)	0.537 ± 0.007	0.531 ± 0.008	0.537 ± 0.003	0.520 ± 0.006	0.528 ± 0.005	0.528 ± 0.004	0.513 ± 0.012	**0.552 ± 0.009**
(Mod, High)	0.451 ± 0.013	0.438 ± 0.008	0.448 ± 0.010	0.434 ± 0.009	0.439 ± 0.008	0.439 ± 0.010	0.442 ± 0.009	**0.467 ± 0.007**
(High, Mod)	0.519 ± 0.008	0.512 ± 0.008	0.514 ± 0.009	0.518 ± 0.005	0.535 ± 0.010	0.489 ± 0.004	0.482 ± 0.008	**0.540 ± 0.008**
(High, High)	0.498 ± 0.014	0.497 ± 0.010	0.495 ± 0.012	0.463 ± 0.008	0.494 ± 0.013	0.483 ± 0.005	0.478 ± 0.016	**0.518 ± 0.003**
Average	0.520 ± 0.010	0.517 ± 0.008	0.519 ± 0.008	0.504 ± 0.007	0.520 ± 0.009	0.515 ± 0.006	0.507 ± 0.011	**0.542 ± 0.007**

RealWorldHAR								
(H, I)	**FedAvg**	**FedProx**	**MOON**	**FedPer**	**FedRoD**	**PmcmFL**	**PEPSY**	**Flux (Ours)**
(Hom, Mod)	0.838 ± 0.002	0.832 ± 0.010	0.836 ± 0.014	0.852 ± 0.016	**0.885 ± 0.003**	0.840 ± 0.006	0.839 ± 0.006	0.867 ± 0.011
(Hom, High)	0.686 ± 0.007	0.673 ± 0.004	0.670 ± 0.005	0.528 ± 0.015	0.704 ± 0.023	0.702 ± 0.013	**0.772 ± 0.009**	0.740 ± 0.004
(Mod, Mod)	0.844 ± 0.006	0.847 ± 0.001	0.844 ± 0.004	0.820 ± 0.015	0.865 ± 0.011	0.842 ± 0.006	0.853 ± 0.005	**0.858 ± 0.007**
(Mod, High)	0.674 ± 0.008	0.673 ± 0.012	0.669 ± 0.008	0.616 ± 0.011	0.693 ± 0.004	0.666 ± 0.005	**0.760 ± 0.017**	0.709 ± 0.006
(High, Mod)	0.838 ± 0.008	0.840 ± 0.003	0.843 ± 0.006	0.835 ± 0.002	0.851 ± 0.006	0.834 ± 0.006	0.803 ± 0.005	**0.855 ± 0.007**
(High, High)	0.709 ± 0.007	0.716 ± 0.010	0.709 ± 0.009	0.720 ± 0.012	0.747 ± 0.009	0.722 ± 0.013	0.710 ± 0.008	**0.766 ± 0.004**
Average	0.765 ± 0.006	0.763 ± 0.007	0.762 ± 0.008	0.729 ± 0.012	0.791 ± 0.009	0.768 ± 0.008	0.789 ± 0.008	**0.799 ± 0.006**

Table 2: Macro-F1 on ADNI with natural intra-client incompleteness under varying inter-client heterogeneity. Extreme assigns each client a unique modality combination.

H	**FedAvg**	**FedProx**	**MOON**	**FedPer**	**FedRoD**	**PmcmFL**	**PEPSY**	**Flux (Ours)**
Homogen.	0.568 ± 0.022	0.570 ± 0.014	**0.572 ± 0.017**	0.511 ± 0.034	0.542 ± 0.019	0.428 ± 0.080	0.509 ± 0.001	0.568 ± 0.012
Moderate	0.511 ± 0.031	0.527 ± 0.020	0.532 ± 0.012	0.505 ± 0.018	0.521 ± 0.022	0.279 ± 0.085	0.492 ± 0.003	**0.533 ± 0.013**
High	0.516 ± 0.037	0.529 ± 0.010	0.530 ± 0.014	0.469 ± 0.026	0.498 ± 0.009	0.203 ± 0.006	0.419 ± 0.050	**0.533 ± 0.020**
Extreme	0.330 ± 0.029	0.350 ± 0.027	0.330 ± 0.024	0.379 ± 0.039	0.389 ± 0.025	0.223 ± 0.048	0.293 ± 0.043	**0.395 ± 0.032**
Average	0.481 ± 0.030	0.494 ± 0.018	0.491 ± 0.017	0.466 ± 0.029	0.487 ± 0.019	0.283 ± 0.055	0.428 ± 0.024	**0.508 ± 0.019**

Models and learning. For health-sensing datasets, we use 1D CNN encoders, a masked attention-based fusion module [2], and a two-layer MLP classifier. FedRoD and `Flux` add a client-private task head to the architecture, while `Flux` additionally includes modality-specific confidence heads during training. We train for 200 global rounds, sampling 30∼50% of clients per round. Each selected client performs three local epochs using SGD. ADNI contains four heterogeneous modalities: MRI, genomic profiles, clinical assessments, and biospecimen measurements. We therefore use modality-specific encoders for these heterogeneous inputs, following prior work [35]. We train ADNI models for 150 global rounds with full client participation and one local epoch per client using Adam. Experiments were conducted using Intel Xeon CPUs and NVIDIA RTX 3090 GPUs.

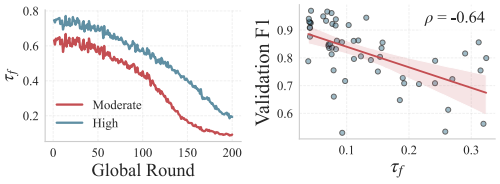

Figure 3: Fused temperature τ_f versus intra-client incompleteness (left) and client-level validation macro-F1 (right).

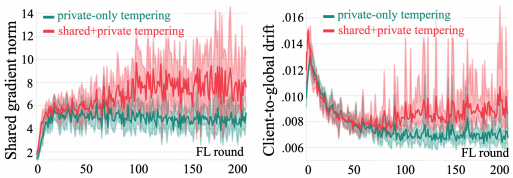

Figure 4: Shared-gradient norm (left) and client-to-global update drift (right) on the PAMAP2 dataset.

Table 3: Calibration performance on PAMAP2 dataset under (Mod, Mod) missingness, averaged over five random seeds.

Metric	FedAvg	FedRoD	PmcmFL	PEPSY	`Flux`
NLL ↓	1.612	**1.366**	1.447	2.165	1.379
ECE ↓	0.193	0.163	0.177	0.220	**0.162**
Brier ↓	0.525	0.500	0.514	0.533	**0.487**

Table 4: Mean fused temperature τ_f under Gaussian input corruption on PAMAP2 with all modalities observed.

FL Round	$\alpha = 0.0$	$\alpha = 0.5$	$\alpha = 0.9$
50	0.300	0.309	0.312
100	0.063	0.075	0.083
150	0.035	0.038	0.041
200	0.029	0.031	0.033

5 Results

Overall Results. As shown in Table 1, `Flux` achieves the highest average macro-F1 across the six dual-axis regimes on each health-sensing dataset: 0.642 on PAMAP2, 0.542 on SleepEDF, and 0.799 on RealWorldHAR. It outperforms all baselines in every regime on PAMAP2 and SleepEDF, and leads overall on RealWorldHAR, including both high inter-client-heterogeneity settings. Table 2 reports results on ADNI, where natural sample-level missingness is retained while client-level modality availability is varied. `Flux` achieves the highest average macro-F1 of 0.508, remains competitive in the homogeneous setting, and performs best under moderate, high, and extreme inter-client heterogeneity. In the extreme setting, where each client has a unique modality composition, `Flux` reaches 0.395, compared with 0.389 for FedRoD. Overall, these results demonstrate consistent performance across varying combinations of sample-level incompleteness and client-level modality heterogeneity.

Confidence and Calibration Diagnostics. Figure 3 examines whether the learned temperature tracks variation in modality completeness and predictive performance. Under moderate inter-client heterogeneity, higher intra-client incompleteness is consistently associated with larger τ_f, indicating lower confidence when fewer modalities are observed. Across all six dual-axis regimes, client-level τ_f is also negatively correlated with validation F1 (Spearman $\rho = -0.64$, $p < 0.001$). Thus, τ_f systematically reflects both modality incompleteness and predictive performance.

We also assess the calibration of the final combined predictions. Although τ_f is used only during training, it shapes private-head optimization and may affect the combined shared and private logits at inference. We evaluate negative log-likelihood (NLL), expected calibration error (ECE), and Brier score on these logits. Table 3 shows favorable calibration performance: `Flux` achieves the best ECE (0.162) and Brier score (0.487), and the second-best NLL (1.379).

Effect of Tempering Placement on Shared Optimization. To isolate the effect of tempering placement, we compare *private-only tempering* (`Flux`) with *shared+private tempering*, which also applies τ_f to the shared task head. Both variants use the same architecture, parameter partitioning, aggregation rule, and hyperparameters on PAMAP2 under the (High, High) dual-axis missingness regime. We measure (i) the mean ℓ_2 norm of gradients in the fusion module and shared task head and (ii) the mean relative client-to-global update drift over all shared parameters. Figure 4 shows that shared+private tempering yields larger and more variable gradient norms and greater update drift, whereas private-only tempering remains lower and more stable after the initial rounds. These results indicate that applying τ_f to both the shared and private heads increases update heterogeneity, while restricting it to private-head training yields a more stable shared optimization trajectory.

Sensitivity to Input Corruption. An observed modality may still be weakly informative or corrupted, which is not captured by a binary mask. To test whether the fused temperature responds to

input quality beyond modality presence, we analyze PAMAP2 in the complete-modality setting with additive Gaussian noise:

$$\widetilde{x} = x + \eta, \qquad \eta \sim \mathcal{N}\big(0, (\alpha\sigma_k)^2\big)\,,$$

where $\alpha \in \{0, 0.5, 0.9\}$ controls the corruption severity and σ_k is the standard deviation of client k's training data. As shown in Table 4, τ_f increases monotonically with the corruption level at every evaluated round (e.g., $0.063 \to 0.075 \to 0.083$ at round 100). This indicates that τ_f responds to input quality rather than modality presence alone, assigning higher temperatures to more corrupted inputs.

Cost Analysis. `Flux` adds lightweight modality-specific confidence heads during training, which are omitted at inference, and a client-private task head. Table 5 compares inference time, training GFLOPs, and macro-F1 on PAMAP2 under the (Mod, Mod) dual-axis missingness regime. `Flux` achieves the highest macro-F1 (0.683) with 19.40 GFLOPs per global round. Although this exceeds FedRoD's 13.97 GFLOPs, it remains well below MOON (34.19) and PEPSY (109.60). Its inference time is 1.44 seconds, compared with 1.16 for FedRoD, 1.71 for PmcmFL, and 1.91 for PEPSY. Overall, `Flux` incurs moderate additional cost over FedRoD for a 1.9-point macro-F1 gain, while remaining substantially cheaper to train than MOON and PEPSY.

Table 5: System cost and macro-F1 on PAMAP2.

Method	Infer. time ↓	Train GFLOPs ↓	F1 score ↑
FedAvg	1.14	11.40	0.634
FedProx	1.09	11.40	0.636
MOON	1.10	34.19	0.636
FedPer	1.11	11.40	0.644
FedRoD	1.16	13.97	0.664
PmcmFL	1.71	11.40	0.643
PEPSY	1.91	109.60	0.641
`Flux`	1.44	19.40	**0.683**

Ablation Study. Table 6 evaluates five ablations of `Flux` on PAMAP2 under high inter-client heterogeneity with moderate and high intra-client incompleteness. *Single-path fused conf.* uses one task pathway and a fusion-level confidence; *Dual-path fused conf.* adds shared-private separation; and *Shared+private temp.* introduces modality-specific confidence heads while tempering both pathways. Two additional variants remove stop-gradient at the fused-representation or unimodal-confidence interface. The complete `Flux` achieves the highest average macro-F1 (0.583) and the best result under (High, High) (0.514). Shared-private separation raises the average from 0.531 to 0.544, modality-specific confidence further improves it to 0.549, and private-only tempering increases it to 0.583, a gain of 0.034. Removing the fused-representation or unimodal stop-gradient reduces the average to 0.575 and 0.569, respectively. Although removing the fused-representation stop-gradient slightly improves (High, Mod) performance (0.653 vs. 0.652), it degrades the more incomplete (High, High) setting from 0.514 to 0.497. These results support modality-specific confidence estimation, private-only tempering, and stop-gradient isolation at both interfaces.

Table 6: Ablation macro-F1 on PAMAP2.

Variant	(High, Mod)	(High, High)	Average
Single-path fused conf.	0.594 ± 0.007	0.468 ± 0.003	0.531 ± 0.005
Dual-path fused conf.	0.621 ± 0.012	0.467 ± 0.012	0.544 ± 0.012
Shared+private temp.	0.616 ± 0.017	0.482 ± 0.007	0.549 ± 0.012
w/o fused-repr. SG	**0.653 ± 0.003**	0.497 ± 0.008	0.575 ± 0.005
w/o unimodal SG	0.638 ± 0.011	0.500 ± 0.022	0.569 ± 0.017
`Flux` (complete)	0.652 ± 0.004	**0.514 ± 0.017**	**0.583 ± 0.011**

6 Conclusion

We studied multimodal federated learning under *dual-axis modality missingness*, where modality availability varies across clients and individual samples may contain only subsets of locally available modalities. We proposed `Flux`, which combines modality-aware confidence tempering with gradient-decoupled private adaptation. The method fuses per-modality confidence estimates into a sample-adaptive training temperature and confines confidence-tempered updates to a client-private pathway, reducing their influence on shared federated optimization. Across diverse health-sensing and biomedical benchmarks, `Flux` achieved the highest average macro-F1 on every dataset. Further analyses show favorable calibration and that the learned temperature responds to modality incompleteness and input corruption, while private-only tempering reduces shared-gradient instability and client-to-global update drift.

References

[1] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[3] Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

[4] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

[5] Qian Dai, Dong Wei, Hong Liu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1445–1453, 2024.

[6] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4035–4045, 2023.

[7] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[8] Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498, 2024.

[9] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–28, 2022.

[10] Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14919–14926, 2023.

[11] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[12] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[13] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[14] Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, 122:103219, 2025.

[15] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.

[16] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[17] 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.

[18] Duong Nguyen, Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. *Advances in Neural Information Processing Systems*, 38:28808–28845, 2025.

[19] Manh Duong Nguyen, Trung Thanh Nguyen, Huy Hieu Pham, Trong Nghia Hoang, Phi Le Nguyen, and Thanh Trung Huynh. Fedmac: Tackling partial-modality missing in federated learning with cross-modal aggregation and contrastive regularization. In *2024 22nd International Symposium on Network Computing and Applications (NCA)*, pages 278–285. IEEE, 2024.

[20] Adiba Orzikulova, Jaehyun Kwak, Jaemin Shin, and Sung-Ju Lee. Federated learning for time-series healthcare sensing with incomplete modalities. *arXiv preprint arXiv:2405.11828*, 2024.

[21] Adiba Orzikulova, Diana A Vasile, Chi Ian Tang, Fahim Kawsar, Sung-Ju Lee, and Chulhong Min. Bioq: Towards context-aware multi-device collaboration with bio-cues. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 504–517, 2025.

[22] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 530–543, 2023.

[23] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296, 2018.

[24] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[25] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE, 2010.

[26] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. Mydj: Sensing food intakes with an attachable on your eyeglass frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[27] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[28] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2016.

[29] Orestis Tsinalis, Paul M Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint arXiv:1610.01683*, 2016.

[30] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE pervasive computing*, 16(4):62–74, 2017.

[31] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. Studentlife: Using smartphones to assess mental health and academic performance of college students. In *Mobile health: Sensors, analytic methods, and applications*, pages 7–33. Springer, 2017.

[32] Michael W Weiner, Paul S Aisen, Clifford R Jack Jr, William J Jagust, John Q Trojanowski, Leslie Shaw, Andrew J Saykin, John C Morris, Nigel Cairns, Laurel A Beckett, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211, 2010.

[33] Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022.

[34] Hyungjun Yoon, Jaehyun Kwak, Biniyam Aschalew Tolera, Gaole Dai, Mo Li, Taesik Gong, Kimin Lee, and Sung-Ju Lee. Selfreplay: Adapting self-supervised sensory models via adaptive meta-task replay. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 226–239, 2025.

[35] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

[36] Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on iot data. In *2022 IEEE/ACM seventh international conference on internet-of-things design and implementation (ioTDI)*, pages 43–54. IEEE, 2022.

A Algorithm

Algorithm 1 Training procedure of `Flux`

Require: Client datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; shared parameters $\theta = \{\{\theta_{E_m}, \theta_{C_m}\}_{m=1}^{M}, \theta_F, \theta_G\}$; private parameters $\{\phi_k\}_{k=1}^{K}$; rounds T; local epochs E

for $t = 0, \ldots, T-1$ **do**
  Select $\mathcal{S}^{(t)}$ and broadcast $\theta^{(t)}$
  for all $k \in \mathcal{S}^{(t)}$ **in parallel do**
    $\theta_k \leftarrow \theta^{(t)}$; load retained ϕ_k
    for $e = 1, \ldots, E$ **do**
      for all minibatches $\mathcal{B} \subset \mathcal{D}_k$ **do**
        $\mathcal{L}^{\mathcal{B}}_{\text{shared}} \leftarrow 0$; $\mathcal{L}^{\mathcal{B}}_{\text{priv}} \leftarrow 0$
        for all $(\mathbf{x}, y, \mathbf{a}) \in \mathcal{B}$ **do**
          $\mathcal{O} \leftarrow \{m : a_m = 1\}$
          for all $m \in \mathcal{O}$ **do**
            $h_m \leftarrow E_m(x_m)$
            $(z_m, \nu_m) \leftarrow C_m(\text{sg}(h_m))$
            $\tau_m \leftarrow \exp(\nu_m/2)$
          end for
          $\mathcal{L}_{\text{uni}} \leftarrow \dfrac{1}{|\mathcal{O}|} \displaystyle\sum_{m \in \mathcal{O}} \text{CE}\left(\frac{z_m}{\tau_m}, y\right)$
          $\tau_f \leftarrow \left(\displaystyle\sum_{m \in \mathcal{O}} \frac{1}{\tau_m^2 + \epsilon} + \epsilon\right)^{-1/2}$
          $h_f \leftarrow F(\{h_m\}_{m \in \mathcal{O}}, \mathbf{a})$
          $z_G \leftarrow G(h_f)$
          $z_{P,k} \leftarrow P_k(\text{sg}(h_f))$
          $\mathcal{L}^{\mathcal{B}}_{\text{shared}} \leftarrow \mathcal{L}^{\mathcal{B}}_{\text{shared}} + \text{CE}(z_G, y) + \mathcal{L}_{\text{uni}}$
          $\mathcal{L}^{\mathcal{B}}_{\text{priv}} \leftarrow \mathcal{L}^{\mathcal{B}}_{\text{priv}} + \text{CE}\left(\dfrac{z_{P,k}}{\text{sg}(\tau_f)}, y\right)$
        end for
        $\mathcal{L}^{\mathcal{B}}_{\text{shared}} \leftarrow \mathcal{L}^{\mathcal{B}}_{\text{shared}}/|\mathcal{B}|$
        $\mathcal{L}^{\mathcal{B}}_{\text{priv}} \leftarrow \mathcal{L}^{\mathcal{B}}_{\text{priv}}/|\mathcal{B}|$
        Update θ_k using $\mathcal{L}^{\mathcal{B}}_{\text{shared}}$
        Update ϕ_k using $\mathcal{L}^{\mathcal{B}}_{\text{priv}}$
      end for
    end for
    Return θ_k; retain updated ϕ_k locally
  end for
  Set each $\theta_m^{(t+1)}$ by aggregating $\{(\theta_{E_m,k}, \theta_{C_m,k}) : k \in \mathcal{S}^{(t)},\ n_{k,m} > 0\}$, weighted by $n_{k,m}$
  Set $(\theta_F^{(t+1)}, \theta_G^{(t+1)})$ by aggregating the corresponding client parameters over $\mathcal{S}^{(t)}$, weighted by n_k
end for

B Construction of Dual-Axis Modality Missingness Regimes

Existing multimodal federated learning studies often model either sample-level dropout or fixed client-level modality subsets [36, 3, 6]. We instead distinguish two coupled axes: *intra-client modality incompleteness*, which determines the modalities observed for each sample, and *inter-client modality heterogeneity*, which determines the modalities available to each client. We instantiate these axes according to the data source. For the three health-sensing datasets, both axes are simulated. For ADNI, we retain naturally missing patient modalities and vary only client-level modality availability.

Health-Sensing Datasets. Health-sensing deployments may exhibit both static differences in device configurations and temporally correlated sensor failures [25]. We model client-level modality availability using a Beta–Bernoulli process. In particular, let $k \in \{1, \ldots, K\}$ index clients, $m \in \{1, \ldots, M\}$ index sensing modalities, and $i \in \{1, \ldots, n_k\}$ index the temporally ordered samples of client k. Here, each modality corresponds to a sensor stream or sensor-location pair defined for the respective dataset. Specifically, each client k draws an availability probability q_k, followed by a binary availability indicator for each modality:

$$q_k \sim \text{Beta}(\alpha_a, \beta_a), \qquad c_{k,m} \mid q_k \sim \text{Bernoulli}(q_k),$$

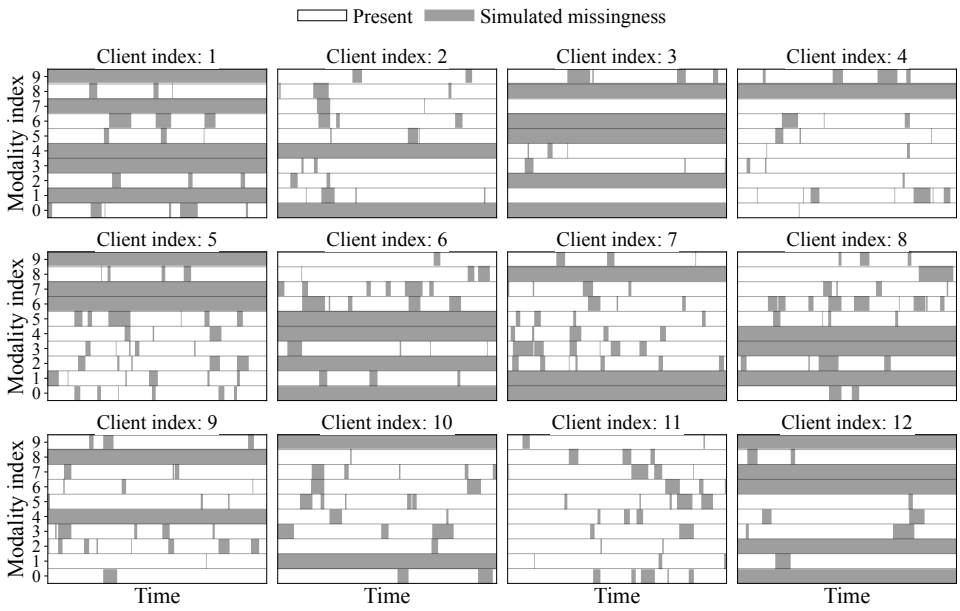

Figure 5: Missingness patterns for 12 RealWorldHAR clients under moderate inter-client heterogeneity and moderate intra-client incompleteness. Clients have different static sensor suites, while available sensors exhibit temporally bursty failures.

where $c_{k,m} = 1$ indicates that modality m belongs to the static sensor suite of client k. Conditional on q_k, the indicators $\{c_{k,m}\}_{m=1}^{M}$ are sampled independently. In the *homogeneous* inter-client setting, we set $c_{k,m} = 1$ for all clients and modalities. The *moderate* and *high* inter-client heterogeneity settings use

$$(\alpha_a, \beta_a) = (45, 20) \quad \text{and} \quad (\alpha_a, \beta_a) = (45, 45),$$

respectively. These configurations correspond to expected modality availability rates of $45/65 \approx 0.692$ and $45/90 = 0.5$. In the homogeneous setting, all clients have the same set of sensor modalities, although sample-level sensor failures may still occur. For each modality available at a client, i.e., $c_{k,m} = 1$, we model its operational state across temporally ordered samples using a two-state Markov chain:

$$s_{k,i,m} \in \{0, 1\}, \qquad \mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix},$$

where

$$p_{uv} = \Pr(s_{k,i,m} = v \mid s_{k,i-1,m} = u), \qquad u, v \in \{0, 1\}.$$

State 0 denotes that the modality is missing, whereas state 1 denotes that it is observed. A large p_{00} produces persistent missing segments, while a large p_{11} produces persistent observed segments, yielding temporally structured failures rather than independent sample-wise dropout. The resulting sample-level observation mask is

$$a_{k,i,m} = c_{k,m} s_{k,i,m}.$$

Modality m is observed for sample i at client k only when it belongs to the client's static modality set and is operational for that particular sample. The client-level variable $c_{k,m}$ therefore defines inter-client modality heterogeneity, whereas $s_{k,i,m}$ defines temporally varying intra-client modality incompleteness.

Figure 5 illustrates the resulting dual-axis patterns. Unlike i.i.d. dynamic dropout [6] or purely static modality removal [3, 36], this construction captures both population-level differences in sensor ownership and within-client temporal sensor instability.

ADNI Dataset. Let k index silos, which serve as FL clients; i index patients assigned to silo k; and $m \in \{1, \ldots, 4\}$ index structural MRI, genomic profiles, clinical assessments, and biospecimen measurements. Let $r_{k,i,m} \in \{0, 1\}$ denote the naturally occurring observation mask. We construct a

five-silo partition using multilabel stratification over diagnosis labels and natural modality masks, producing similar label and missingness-pattern distributions across silos. In the *homogeneous* setting, no modality is removed at the silo level:

$$c_{k,m} = 1, \qquad a_{k,i,m} = r_{k,i,m}.$$

For the *moderate* and *high* inter-client heterogeneity settings, we retain the same patient-to-silo assignment and sample a static modality profile for each silo:

$$p_{a,k} \sim \mathrm{Beta}(\alpha_a, \beta_a), \qquad c_{k,m} \mid p_{a,k} \sim \mathrm{Bernoulli}(p_{a,k}),$$
$$a_{k,i,m} = r_{k,i,m} c_{k,m}.$$

We use $(\alpha_a, \beta_a) = (17, 3)$ for moderate heterogeneity and $(6, 14)$ for high heterogeneity. If a silo's sampled modality-availability vector is all-zero, we set one randomly chosen modality to available so each silo retains at least one modality. Patients with $\sum_{m=1}^{M} a_{k,i,m} = 0$ after applying simulated missingness to the natural observation mask are excluded. The *extreme* setting is constructed without stochastic modality removal. Instead, patients are grouped by their exact nonempty natural modality mask, such that each client has a unique modality combination.

C Related Work

Prior multimodal FL methods address different missingness regimes. PmcmFL uses prototype masking and contrastive learning for within-client missing modalities, while MFCPL applies complete prototypes and cross-modal regularization under severe missingness [3, 14]. FedMAC handles complete and partial modality missingness through imputation embeddings, and contrastive regularization [19]. Harmony and FedMEMA focus on heterogeneous modality sets across clients using modality-wise training or modality-specific encoders and multimodal anchors [22, 5], while PEPSY addresses cross-client and within-client missing-data patterns through representation reconfiguration [18]. These methods primarily compensate for missing representations or coordinate learning across modality configurations. Personalized FL motivates shared-private parameterization [1, 4], while input-dependent uncertainty modeling and adaptive temperature scaling support evidence-dependent confidence control [13, 10]. In contrast, `Flux` models the reliability of each observed modality during federated training and fuses these estimates into a sample-adaptive temperature that reflects both evidence completeness and quality. It then confines confidence-tempered adaptation to a gradient-decoupled private pathway, preventing evidence-dependent tempering from modifying shared task gradients. Thus, `Flux` addresses the optimization tension between sample-specific evidence adaptation and stable cross-client representation learning.

D Experiment Details

D.1 Datasets

We use four real-world multimodal datasets: three health-sensing benchmarks (PAMAP2 [24], RealWorldHAR [28], and SleepEDF [7, 12]) and the biomedical cohort ADNI [32], which contains naturally missing data.

PAMAP2 [24] consists of recordings from nine users performing twelve activities using wearable Inertial Measurement Unit (IMU) sensors. Following prior work [9], we exclude one subject who contributed data for only a single activity, resulting in eight clients. The dataset provides accelerometer and gyroscope signals from three body locations: wrist, chest, and ankle, yielding six distinct sensing modalities in total.

SleepEDF [7, 12] contains sleep recordings from 20 participants, including electroencephalography (EEG), electrooculography (EOG), chin electromyography (EMG), respiration signals, and event markers. Each recording is annotated with hypnograms containing five sleep stages. Following prior work [29, 23], we utilize the Sleep Cassette subset, which focuses on age-related sleep patterns in healthy individuals and is commonly used for sleep-stage classification.

RealWorldHAR [28] consists of activity recordings from fifteen participants performing eight daily activities. Data were collected with seven body-worn IMU sensors, two of which were discarded due

Table 7: Hyperparameter configurations.

Method	HP	PAMAP2	RealWorldHAR	SleepEDF	ADNI
FedAvg	Learning Rate	0.001	0.03	0.03	0.0001
FedPer	Learning Rate	0.001	0.03	0.001	0.0001
FedProx	Learning Rate	0.001	0.03	0.01	0.0001
	Proximal Term	0.1	0.01	0.01	0.01
MOON	Learning Rate	0.001	0.03	0.03	0.0001
	Contrast. Weight	10	0.1	10	0.1
	Temperature	0.5	0.5	1.0	1.0
PmcmFL	Learning Rate	0.001	0.03	0.001	0.0001
	CLIP Loss Weight	0.1	0.01	0.5	0.5
FedRoD	Learning Rate	0.001	0.03	0.01	0.0001
PEPSY	Learning Rate	0.01	0.01	0.001	0.00004
`Flux`	Learning Rate	0.001	0.001	0.001	0.0001
	Conf. Head Hidden Dim.	512	512	512	128

to insufficient activity coverage. The final dataset comprises signals from ten modalities, spanning five body locations and two IMU sensor types.

ADNI [32] is a multimodal study of Alzheimer's disease and aging. Each patient may contribute up to four data types: structural MRI, genomic profiles, clinical assessments, and biospecimen measurements. We cast the task as three-way diagnostic classification among cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Following our federated evaluation protocol, participants are partitioned into five silos that serve as FL clients. Unlike the wearable benchmarks, modality incompleteness in ADNI is *natural*, meaning that not all patients have all four modalities available. In our experiments, we retain the naturally occurring intra-client modality incompleteness and vary only inter-client modality heterogeneity by changing modality availability across silos.

D.2 Baselines

FedAvg [17] is the standard federated learning baseline that enables collaborative training without sharing raw data but provides no explicit mechanism for handling missing modalities.

FedProx [16] addresses system and statistical heterogeneity. It enhances performance by adding a proximal term to the local training loss, penalizing deviations between local and global models to improve stability and convergence.

MOON [15] targets the problem of local data heterogeneity. It incorporates contrastive learning into federated learning, encouraging alignment between the global and local models' embeddings while pushing apart embeddings from the client's previous local model.

FedPer [1] addresses statistical heterogeneity by splitting models into shared base layers and client-specific personalization layers. The base layers are trained collaboratively across clients using FedAvg, while the personalization layers are updated only with local data.

FedRoD [4] bridges generic and personalized federated learning through a decoupled prediction architecture. It decouples the local model into two predictors: a generic head trained with balanced risk minimization to improve robustness against non-IID class distributions, and a personalized head trained with empirical risk minimization to capture client-specific patterns.

PmcmFL [3] introduces a prototype library to address the challenges of missing modalities in federated multimodal learning. Prototypes are used both as masks for absent modalities and as anchors in a contrastive loss to reduce client heterogeneity.

PEPSY [18] tackles the multimodal FL regime where clients have different modality subsets and also suffer within-modality missing features. It learns client-side data-missing profiles to adapt and align

the shared representation to each client's missingness pattern, and aggregates these signals across clients with similar patterns.

D.3 Training and Evaluation Details

Hyperparameter tuning. Table 7 reports the final selected hyperparameters for each method and dataset. For the sensing datasets, we tune the learning rate over $\{0.001, 0.01, 0.03, 0.05\}$ and select the value with the best validation performance. For ADNI, we use an ADNI-specific grid consisting of local learning rates in $\{4 \times 10^{-5}, 10^{-4}\}$ and batch sizes in $\{8, 16\}$, due to its larger input dimensionality and memory footprint.

We additionally tune the FedProx proximal coefficient μ_{prox} over $\{0.001, 0.01, 0.1, 0.5, 1\}$ for the sensing datasets and $\{0.01, 0.1\}$ for ADNI. For MOON, we tune the contrastive weight μ_{contrast} over $\{0.1, 1, 5, 10\}$ and the temperature τ over $\{0.1, 0.5, 1\}$ for the sensing datasets; for ADNI, the corresponding grids are $\{0.1, 10\}$ and $\{0.5, 1.0\}$. For PmcmFL, we tune the CLIP loss weight over $\{0.01, 0.1, 0.5, 1.0, 5.0\}$ for the sensing datasets and $\{0.01, 0.1, 0.5\}$ for ADNI. For `Flux`, we tune the confidence-head hidden dimension over $\{128, 256, 512\}$ on all datasets.

Model selection. For global methods (FedAvg, FedProx, MOON, PmcmFL, and PEPSY) the server selects the checkpoint with the highest mean validation macro-F1 across clients. For personalized methods (FedPer, FedRoD, and `Flux`) each client independently selects the checkpoint with the highest local validation macro-F1. Final test performance is computed using the selected global or client-specific checkpoints, respectively.

Evaluation consistency and hardware. For each dataset, missingness regime, and random seed, all methods used the same train, validation, and test partitions and the same generated client- and sample-level modality masks. Experiments were conducted using Intel Xeon CPUs and NVIDIA RTX 3090 GPUs. The server was equipped with 432 GB of RAM and ran Ubuntu 22.04.2 LTS with CUDA 11.5.