LABEL-PRIOR LEAKAGE IN FEDERATED EVALUATION AND AN INFERENCE MANAGEMENT FIX

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated learning enables collaborative medical AI without sharing patient data, yet post-deployment inference and evaluation remain largely unmanaged. We identify a previously overlooked label-prior leakage in personalized federated evaluation: client-matched testing of private heads implicitly reveals client identity and its label distribution, inflating apparent performance even for a pure prior-only predictor. To address both training heterogeneity and reliable deployment, we introduce FedGIM, a proximal-regularized method with private heads and class-balanced local objectives, together with Federated Inference Management (FedIM). FedIM is a model-agnostic post-training layer that coordinates four stages: Adapt (entropy-minimizing test-time adaptation), Monitor (MMD and Kolmogorov-Smirnov drift detection), Act (confidence-gated computational escalation, optional differential-privacy logit release, and model fusion), and Refer (selective classification with explicit coverage control and human referral). Evaluated on non-IID endoscopic screening under a leakage-aware pooled protocol with label-free feature-space routing, FedGIM recovers near-centralized accuracy while FedIM provides controllable coverage-risk trade-offs, drift response, latency-accuracy balance, and privacy-utility curves. Our results demonstrate that evaluation protocol choice can dominate reported gains and that managed inference is essential for clinically trustworthy federated systems.

1 INTRODUCTION

Federated learning (FL) enables collaborative medical AI without sharing patient data by exchanging model updates across institutions McMahan et al. (2017); Guan et al. (2024). Heterogeneity in populations, protocols, and equipment produces non-IID data, motivating extensive work on robust aggregation, proximal terms, and personalization Li et al. (2020); Ng et al. (2023); Hernandez-Cruz et al. (2024), and training under such conditions remains a core challenge Guan et al. (2024); Kaissis et al. (2021).

FL addresses training, not post-deployment behavior. After convergence, inference is typically a fixed forward pass (Fig. 2), yet clinical deployment involves acquisition shifts, resource limits, uncertain predictions, and cases requiring human oversight Ng et al. (2023); Guan et al. (2024), with distribution shift, uncertainty quantification, and clinical utility remaining open problems Liang et al. (2024).

Existing remedies operate largely in isolation: test-time entropy minimization adapts models at inference Wang et al. (2021b), maximum mean discrepancy (MMD) detects distribution drift Gretton et al. (2012), dynamic networks allocate computation by input difficulty Han et al. (2021), and selective classification formalizes abstention under uncertainty Geifman & El-Yaniv (2017). These techniques enable adaptive, efficient, uncertainty-aware inference, but no unified decision framework coordinates them.

We therefore introduce Federated Learning Inference Management (FedIM), a post-training layer that treats deployment as a managed process with four coordinated functions: Adapt (lightweight test-time adaptation), Monitor (drift and confidence assessment), Act (computational escalation, privacy-aware release, selective prediction), and Refer (abstention and human review). As shown in Fig. 2, FedIM operates on personalized models from federated aggregation and remains model-agnostic.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

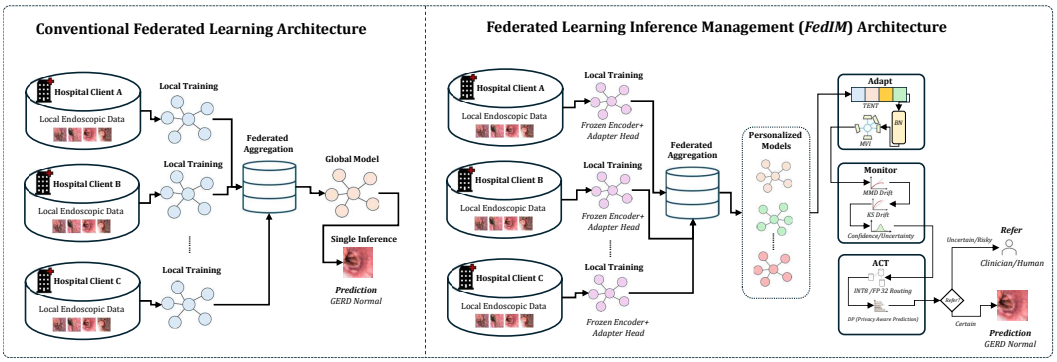


Figure 1: Comparison of conventional federated learning (left) with fixed global-model inference versus the proposed FedIM architecture (right), which adds Adapt, Monitor, Act, and Refer stages around personalized models for managed clinical deployment.

We evaluate FedIM on non-IID endoscopic screening with FedGIM (proximal regularization, private head, class-balanced optimization) against a FedAvg baseline McMahan et al. (2017), covering accuracy, latency, drift response, coverage-risk trade-offs, escalation, and privacy-utility behavior. We further identify label-prior leakage in client-matched personalized evaluation and propose a pooled held-out protocol with label-free feature-space routing Tan et al. (2022). Our contributions are: (1) a model-agnostic FedIM layer coordinating adaptation, monitoring, action, and referral; (2) an operational evaluation framework spanning performance, cost, uncertainty, and referral; and (3) a leakage-aware protocol for personalized FL. Together they shift emphasis from training alone toward reliable, managed inference.

2 Related Work

Heterogeneity and medical imaging. FedAvg established communication-efficient decentralized training as the basis for federated learning McMahan et al. (2017). FedProx added a proximal term constraining local updates to the global model, improving convergence under statistical heterogeneity Li et al. (2020), and empirical studies confirm it reduces client drift at the cost of an extra hyperparameter Kim et al. (2021). Surveys of non-IID federated learning categorize heterogeneity into label, quantity, and feature skew, and review mitigations such as weighted aggregation, batch-norm averaging, and control-variate corrections (SCAFFOLD) Ma et al. (2023). In medical imaging, non-IID data across scanners, protocols, and populations remains a persistent bottleneck Guan et al. (2024); batch-norm averaging and entropy-based priors reduce activation divergence, while disentanglement methods separate anatomical shape from appearance to improve on baseline federated learning Ng et al. (2023); Hernandez-Cruz et al. (2024). End-to-end privacy-preserving federated learning for multi-institutional imaging underscores the joint need for privacy and robustness in clinical pipelines Kaissis et al. (2021).

Personalization and evaluation leakage. Personalization approaches either adapt a global model or maintain client-specific heads Tan et al. (2022). Many such methods conflate client identity with the test distribution; newer proximal-plus-head designs keep heads unshared so personalization is not overwritten during aggregation Wang et al. (2024). This supports our claim that client-matched evaluation leaks label-prior information, motivating pooled, label-free routing.

Test-time adaptation and drift. TENT adapts batch-normalization affine parameters via entropy minimization without labels Wang et al. (2021b). AdaBN and prediction-time correction show that replacing source with target-batch statistics improves robustness under covariate shift given sufficient batch size Li et al. (2016); Schneider et al. (2020), and updating only affine parameters is more stable in the low-sample regimes typical of clinical inference Wu et al. (2023); Yang et al. (2022). Continual test-time adaptation surveys highlight catastrophic forgetting and error accumulation as open risks over long deployment horizons Liang et al. (2024); Maharana et al. (2025). Dataset shift

is the rule rather than the exception in deployed medical AI Finlayson et al. (2021), motivating shift-stable model design via causal structure Subbaswamy & Saria (2020). Gretton et al.'s kernel two-sample test (MMD) remains the standard nonparametric detector underlying most monitoring pipelines Gretton et al. (2012), extended to automatic shift attribution and remediation in production Roschewitz et al. (2024); Subasri et al. (2025), with Quinonero-Candela et al. providing the canonical taxonomy of covariate, label, and concept shift Quiñonero-Candela et al. (2009).

Selective prediction and adaptive computation. Selective classification is formalized through a risk-coverage trade-off underlying referral-based abstention Geifman & El-Yaniv (2017). Abstention thresholds can be calibrated against predictive-value requirements without retraining Ko et al. (2026), though abstention can worsen subgroup disparities when uncertainty concentrates in under-represented populations Sargeant et al. (2025). Surveys of dynamic neural networks and early-exit/split-computing architectures support input-dependent inference cost Han et al. (2021); Matsubara et al. (2022); Laskaridis et al. (2021), motivating precision-routing schemes (e.g., INT8 vs. FP32) based on input difficulty.

Privacy. Differentially private federated learning generally preserves clinical utility at moderate budgets, with strict budgets causing accuracy loss amplified in minority subgroups Kim et al. (2021); Mohammadi et al. (2026); Shen et al. (2022). Deep Leakage from Gradients and its successors (iDLG, R-GAP) show gradient updates alone can reconstruct training images, motivating privacy-aware release as a necessary complement to federated learning's structural privacy Zhu et al. (2019); Zhao et al. (2020); Geiping et al. (2020); Zhu & Blaschko (2021); Gong et al. (2023).

Endoscopic GERD classification. GERD-VGGNet and its successor GerdNet-RF exceed trainee-endoscopist accuracy in Los Angeles grade classification, establishing a baseline for single-inference models without inference-time management Wang et al. (2021a); Yen et al. (2022); Chan et al. (2023); Ge et al. (2023).

3 METHODOLOGY

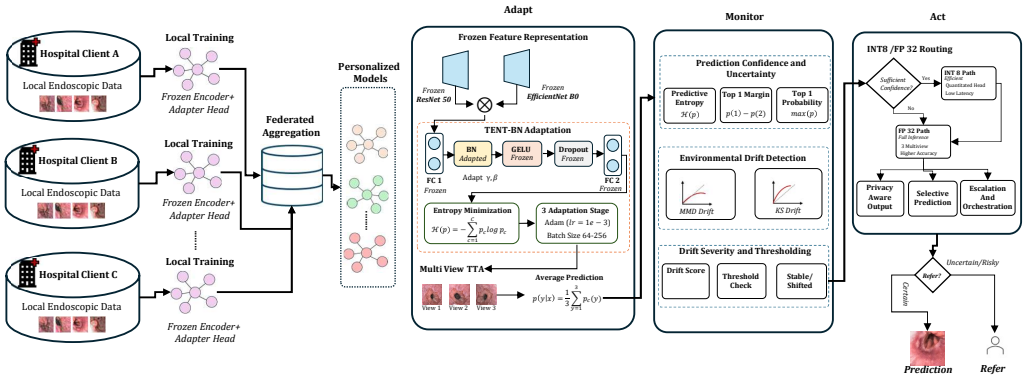


Figure 2: Overview of FedGIM training and the FedIM inference-management layer (Adapt, Monitor, Act, Refer).

3.1 PROBLEM FORMULATION AND SYSTEM OVERVIEW

Let K clients hold local datasets $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n_k}$, $y_i \in \{1, \ldots, C\}$, drawn from client-specific distributions $\mathcal{P}_k$ with $\mathcal{P}_k \neq \mathcal{P}_j$ (non-IID). A frozen feature extractor $\phi_{\text{enc}} : \mathcal{X} \to \mathbb{R}^d$ is shared across clients, and each client trains a lightweight head $f_{\theta,\varphi_k} : \mathbb{R}^d \to \Delta^{C-1}$ decomposed into a shared trunk θ and a private component φ_k. Federated training (Sec. 3.2) yields $(\theta^\star, \{\varphi_k^\star\})$. FedIM (Sec. 3.3) is a deployment-time operator

$$\mathcal{M}_{\text{IM}} : (x, \mathcal{S}_t) \longmapsto (\hat{y}, \text{acc} \in \{0, 1\}), \tag{1}$$

mapping an input x and a running statistics state $\mathcal{S}_t$ to a released prediction $\hat{y}$ and an accept/refer decision, without altering $(\theta^\star, \varphi_k^\star)$.

3.2 Federated Model Development with FedGIM

FedGIM combines three complementary mechanisms for non-IID federated learning: a class-balanced local objective, a FedProx-style proximal constraint (Li et al., 2020), and a FedPer-style private prediction head. For client k, let $n_{k,c}$ denote the number of local samples belonging to class c. We use effective-number weights

$$w_{k,c} = \frac{1-\beta}{1-\beta^{n_{k,c}}}, \qquad \beta \in [0,1), \tag{2}$$

with normalization across classes. The resulting class-balanced loss is combined with a proximal penalty on the shared parameters:

$$\min_{\theta,\{\varphi_k\}} \sum_{k=1}^{K} \frac{n_k}{n} \left[\mathcal{L}_k(\theta, \varphi_k) + \frac{\mu}{2} \left\| \theta - \theta^{(t)} \right\|_2^2 \right], \qquad n = \sum_{k=1}^{K} n_k. \tag{3}$$

After local optimization, only the shared parameters are aggregated:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)}, \qquad \varphi_k^{(t+1)} \text{ retained locally.} \tag{4}$$

Thus, the shared representation is globally coordinated while the client-specific head remains personalized. Setting $\mu = 0$ and removing the private head recovers the FedAvg setting (McMahan et al., 2017), while the proximal and private-head components correspond to FedProx and FedPer-style personalization, respectively. The detailed local objective, weight normalization, and optimization procedure are provided in Appendix A.

Leakage-aware test-time routing. Personalized FL evaluation can become confounded when a personalized model is evaluated on the held-out shard of the same client whose identity determines the selected head. Under non-IID class distributions, this procedure can implicitly expose client-specific label priors to the prediction system. We therefore evaluate all methods on a pooled held-out test set and select personalized heads using a label-free feature-space router:

$$r(x) = \arg \max_{k \in \{1,\dots,K\}} \cos(\phi_{\text{enc}}(x), \mu_k), \qquad \mu_k = \frac{\sum_{x_i \in \mathcal{D}_k^{\text{train}}} \phi_{\text{enc}}(x_i)}{\left\| \sum_{x_i \in \mathcal{D}_k^{\text{train}}} \phi_{\text{enc}}(x_i) \right\|_2}. \tag{5}$$

The centroids are computed exclusively from training data, so no test labels or true test-client identities are available during routing. The resulting protocol and the controlled prior-only leakage analysis are described in Appendix H.2.

3.3 Inference Management Framework

FedIM treats inference as a managed decision process rather than a single forward pass. Given an input batch, the system first monitors predictive reliability and distributional change, adapts the lightweight prediction path when required, selects an appropriate computational and privacy operating point, and finally determines whether each prediction should be accepted or referred. The four stages are therefore organized as

$$\text{Monitor} \rightarrow \text{Adapt} \rightarrow \text{Act} \rightarrow \text{Refer}. \tag{6}$$

The detailed implementation of each stage is provided in Appendix C.

Monitor. The Monitor stage evaluates both prediction reliability and distributional change. Predictive confidence, entropy, and top-1 margin characterize the current prediction state, while complementary feature-distribution tests, including maximum mean discrepancy (MMD) and the Kolmogorov–Smirnov (KS) test, detect acquisition or representation shift. The resulting drift signal controls whether test-time adaptation is activated.

Algorithm 1 FedIM inference policy

Require: input batch $\mathcal{B}$, reference window $\mathcal{R}$, models $\{f_{\theta,\varphi_k}\}$, thresholds $(\tau_{\mathrm{MMD}}, \alpha, \tau_{\mathrm{conf}}, \tau_{\mathcal{H}}, \tau_m, \rho, \epsilon, \delta)$

$Z \leftarrow \phi_{\mathrm{enc}}(\mathcal{B})$
$D \leftarrow$ drift indicator from Eq. equation 35 on $(\mathcal{R}, Z)$ ▷ Monitor
$a \leftarrow \mathbb{1}[\text{TTA enabled}] \vee D$ ▷ Eq. equation 40
if $a = 1$ **then**
    update BatchNorm parameters (γ, β) on $\mathcal{B}$ using Eq. equation 39 ▷ Adapt
end if
for each $x \in \mathcal{B}$ **do**
    $k \leftarrow r(x)$ via Eq. equation 5
    $\bar{p}^{(\mathrm{cheap})}(x) \leftarrow Q(f_{\theta,\varphi_k})(x)$
    $e(x) \leftarrow$ Eq. equation 42
    if $e(x) = 1$ **then**
        compute $\bar{p}^{(\mathrm{full})}(x)$ over V augmented views
    end if
    $p(x) \leftarrow$ Eq. equation 43
    if global model is available **then**
        $p_{\mathrm{out}}(x) \leftarrow p_{\mathrm{fused}}(x)$ via Eq. equation 44
    else
        $p_{\mathrm{out}}(x) \leftarrow p(x)$
    end if
    $\tilde{z}(x) \leftarrow \log p_{\mathrm{out}}(x) + \mathcal{N}(0, \sigma^2 I)$ via Eq. equation 46 ▷ Act
end for
$\tilde{p}(x) \leftarrow \mathrm{softmax}(\tilde{z}(x))$
$\mathrm{acc}(\cdot) \leftarrow$ Eq. equation 53 over $\mathcal{B}$ using coverage ρ ▷ Refer
return $\{(\hat{y}(x) = \arg\max_c \tilde{p}_c(x), \mathrm{acc}(x))\}_{x \in \mathcal{B}}$

Adapt. When enabled or triggered by detected drift, FedIM performs label-free test-time adaptation following TENT (Wang et al., 2021b). Only the BatchNorm affine parameters of the lightweight prediction path are updated through entropy minimization, while the visual encoders and federated prediction parameters remain frozen. Multiple augmented views can subsequently be aggregated to stabilize the adapted prediction. The detailed TENT objective, parameter update, gating mechanism, and multi-view procedure are given in Appendix C.2.

Act. The Act stage controls computational escalation, model fusion, and inference-time privacy. Each input is first evaluated through a low-cost quantized head. Cases below the confidence threshold are escalated to the full-precision multi-view path. When both personalized and global models are available, their predictions are fused using a fixed weight. The resulting logits can then be perturbed before release using a Gaussian mechanism. These operations are detailed in Appendix C.3.

Refer. The final stage converts the released prediction into an operational accept-or-refer decision. Normalized entropy and the top-1 margin jointly characterize uncertainty, while a calibration-derived coverage threshold controls the fraction of cases eligible for automated acceptance. Cases outside the permitted uncertainty region are referred for human review. The complete selective-prediction formulation is provided in Appendix C.4.

Ablating one component at a time while holding the remaining stages fixed yields the variant family

$$\mathcal{V} = \{\text{Full-IM}\} \cup \{\text{w/o } c : c \in \{\text{Adapt}, \text{Monitor}, \text{Escalation}, \text{Privacy}, \text{Refer}\}\} \cup \{\text{No-IM}\}. \quad (7)$$

The No-IM reference uses the full-precision prediction directly, with $D \equiv 0$, $e \equiv 1$, $\sigma \equiv 0$, and $\mathrm{acc} \equiv 1$.

3.4 Privacy, Deployment, and Operational Constraints

Since $\tilde{z}$ in Eq. equation 46 is the only quantity released, the privacy guarantee is stated at the output level. For neighboring inputs x, x' satisfying $\|z(x) - z(x')\|_2 \leq \Delta$, the Gaussian mechanism

satisfies

$$\Pr[\tilde{z}(x) \in S] \leq e^{\epsilon} \Pr[\tilde{z}(x') \in S] + \delta, \qquad \forall S. \tag{8}$$

This output-level guarantee is independent of the training-time privacy of $(\theta^\star, \varphi_k^\star)$, which is not separately privatized here.

The computational cost of FedIM is characterized by the escalation rate

$$\bar{e} = \mathbb{E}[e(x)] \tag{9}$$

and the mixed per-item latency

$$\ell(x) = \ell_{\text{enc}} + \ell_{\text{cheap}} + e(x)\ell_{\text{full}} + \ell_{\text{drift}}. \tag{10}$$

Thus, throughput is determined by $1/\mathbb{E}[\ell(x)]$ under the realized mixture of inexpensive and escalated inference paths rather than by either path in isolation.

3.5 Evaluation Protocol

Binary screening performance is summarized at a fixed operating specificity s_0 using

$$\text{Sens@}s_0 = \text{TPR}(t^\star), \qquad t^\star = \inf\{t : \text{TNR}(t) \geq s_0\}. \tag{11}$$

We additionally report AUC-ROC and expected calibration error:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{I}_b|}{N} \left|\text{acc}(\mathcal{I}_b) - \text{conf}(\mathcal{I}_b)\right|. \tag{12}$$

Because the router in Eq. equation 5 places all personalized methods on the same pooled test set, paired AUC comparisons use DeLong's test. For AUC estimates $\hat{A}_1$ and $\hat{A}_2$, with covariance matrix Σ,

$$z = \frac{\hat{A}_1 - \hat{A}_2}{\sqrt{\Sigma_{11} + \Sigma_{22} - 2\Sigma_{12}}}, \qquad p = 2\Phi(-|z|). \tag{13}$$

Across-seed robustness is assessed using the Wilcoxon signed-rank statistic with Holm–Bonferroni family-wise correction and effect size

$$d_z = \frac{\bar{\Delta}}{s_\Delta} \tag{14}$$

over $S = 5$ seeds. These seed-level analyses are treated as secondary given the corresponding finite resolution of paired significance testing.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed federated training and inference-management framework under non-IID endoscopic screening conditions. The federated training component is denoted as FedGIM, while FedIM denotes the model-agnostic inference management layer applied after federated training. Unless otherwise stated, results are reported as mean $\pm$ standard deviation over five random seeds. All methods are evaluated on the same held-out test examples under the leakage-aware pooled protocol described in Section 3.5. We report AUROC, sensitivity at 90% specificity, F1, accuracy, expected calibration error (ECE), coverage, latency, and human-in-the-loop (HITL) referral rate according to the requirements of each experiment.

4.2 Main Federated Performance

We first compare federated training strategies under the leakage-aware held-out evaluation protocol. Table 1 summarizes the performance of centralized training, local-only training, standard federated baselines, and the proposed FedGIM configuration. The results show that the federated approaches recover much of the centralized performance while preserving the decentralized

Table 1: Federated learning performance under the leakage-aware held-out protocol. Results are reported as mean ± std over five seeds.

Method	AUROC ↑	Sens. @ 90% Spec. ↑	Specificity ↑	F1 ↑	Accuracy ↑
Centralized†	0.8599 ± 0.0020	0.6225 ± 0.0134	0.9012 ± 0.0041	0.7284 ± 0.0087	0.7648 ± 0.0069
Local-only	0.7707 ± 0.0068	0.4175 ± 0.0363	0.8724 ± 0.0115	0.6128 ± 0.0194	0.6642 ± 0.0173
FedAvg	0.8429 ± 0.0178	0.5901 ± 0.0449	0.8943 ± 0.0098	0.7015 ± 0.0213	0.7491 ± 0.0215
FedProx	0.8447 ± 0.0151	0.5888 ± 0.0461	0.8951 ± 0.0102	0.7032 ± 0.0207	0.7498 ± 0.0226
FedPer	0.7793 ± 0.0299	0.4369 ± 0.0651	0.8789 ± 0.0147	0.6284 ± 0.0281	0.6735 ± 0.0329
pFedMe	0.7603 ± 0.0215	0.4085 ± 0.0455	0.8697 ± 0.0132	0.6051 ± 0.0229	0.6593 ± 0.0224
FedGIM	$\mathbf{0.8512 \pm 0.0124}$	$\mathbf{0.6087 \pm 0.0318}$	$\mathbf{0.8985 \pm 0.0076}$	$\mathbf{0.7159 \pm 0.0152}$	$\mathbf{0.7573 \pm 0.0148}$

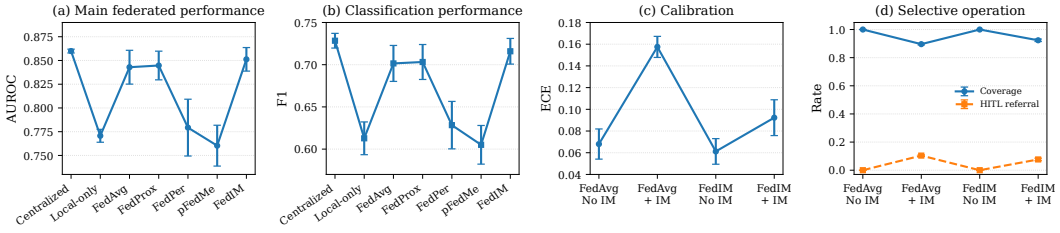


Figure 3: Main predictive performance under the leakage-aware held-out evaluation protocol. Exact numerical results are reported in Table 1.

training setting. Among the evaluated federated configurations, FedGIM obtains an AUROC of 0.8512 ± 0.0124, sensitivity of 0.6087 ± 0.0318 at 90% specificity, F1 of 0.7159 ± 0.0152, and accuracy of 0.7573 ± 0.0148. Figure 3 provides a compact visualization of the main predictive results. The figure emphasizes the relative behavior of the federated methods across AUROC and F1 while the table retains the exact numerical results.

4.3 Effect of Inference Management

We next examine whether the proposed FedIM layer provides consistent inference-time behavior when attached to different federated learners. Table 2 reports the corresponding predictive, calibration, coverage, and HITL referral metrics. For FedGIM, attaching FedIM changes AUROC from 0.8572 ± 0.0154 to 0.8419 ± 0.0142 and F1 from 0.7218 ± 0.0287 to 0.6984 ± 0.0256, while ECE changes from 0.0612 ± 0.0118 to 0.0923 ± 0.0165. At the same time, the managed configuration introduces explicit coverage control and HITL referral, which are absent from the raw inference configuration. These results illustrate that FedIM should be evaluated not only through predictive metrics but also through its calibration and selective-decision behavior. A privacy-utility and covariate-drift stress test of this same layer is reported in Appendix H.1.

4.4 Component Ablation and Operational Behavior

We perform a one-component-at-a-time ablation to quantify the contribution of the individual FedIM components. The full configuration obtains an AUROC of 0.8518 ± 0.0126, F1 of 0.7124 ± 0.0248, and ECE of 0.0589 ± 0.0157. It also reaches a coverage of 0.9287 ± 0.0093 with a HITL referral rate of 0.0713 ± 0.0093. The ablation results indicate that removing individual inference-management components changes different aspects of the deployment behavior. In particular, removing privacy, TTA, drift detection, orchestration, or HITL changes the predictive, calibration, coverage, or referral characteristics of the system. Figure 4 provides a compact view of the component-level effects.

4.5 Selective Prediction and Human Referral

We evaluate selective prediction by varying the retained coverage. Figure 5 presents the resulting risk–coverage behavior for FedAvg and FedGIM with inference management. At 50% coverage, selective error is 0.2248 ± 0.0126 for FedAvg and 0.1217 ± 0.0210 for FedGIM with FedIM. At 100% coverage, the corresponding errors are 0.3077 ± 0.0165 and 0.2436 ± 0.0211. The selective false-negative rate follows the same pattern. At 50% coverage, it is 0.2955 ± 0.0297 for FedAvg compared

Table 2: Effect of inference management when attached to different federated learners. Results are reported as mean ± std over five seeds.

Training model	IM	AUROC ↑	F1 ↑	ECE ↓	Coverage ↑
FedAvg	No	0.8495 ± 0.0180	0.7095 ± 0.0332	0.0680 ± 0.0139	1.0000 ± 0.0000
FedAvg	Yes	0.8184 ± 0.0201	0.6395 ± 0.0391	0.1575 ± 0.0097	0.8960 ± 0.0053
FedGIM	No	$\mathbf{0.8572 \pm 0.0154}$	$\mathbf{0.7218 \pm 0.0287}$	$\mathbf{0.0612 \pm 0.0118}$	1.0000 ± 0.0000
FedGIM	Yes	0.8419 ± 0.0142	0.6984 ± 0.0256	0.0923 ± 0.0165	$\mathbf{0.9235 \pm 0.0087}$

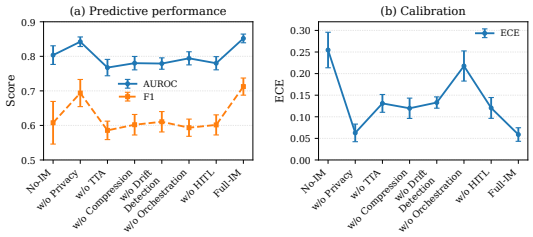


Figure 4: Component-level ablation of FedIM. The heatmap summarizes the effect of removing individual inference-management components on AUROC, F1, ECE, coverage, and HITL referral rate.

with 0.1086 ± 0.0211 for FedGIM with FedIM. These results demonstrate the coverage-dependent operating behavior of the managed inference system. As detailed in Section 3.5 and validated in Appendix H.2, all results above use a pooled, label-free evaluation protocol; client-matched evaluation was found to substantially inflate apparent performance through label-prior leakage.

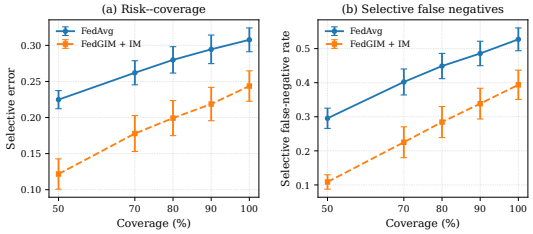


Figure 5: Selective prediction under varying retained coverage. FedGIM with FedIM is compared with FedAvg using selective error and selective false-negative rate.

5 Conclusion

We have shown that client-matched evaluation of personalized federated models systematically leaks label-prior information, producing large artifactual gains that disappear under a pooled label-free routing protocol. FedGIM combines proximal regularization of a shared trunk with private heads and class-balanced losses to improve non-IID endoscopic screening performance, while the model-agnostic FedIM layer turns fixed inference into a managed process of adaptation, drift monitoring, computational and privacy-aware action, and selective referral. Experiments confirm that FedIM supplies explicit coverage control, drift responsiveness, and tunable privacy-utility behavior without altering the underlying trained models. These findings shift the emphasis of federated medical AI from training alone toward evaluation integrity and operational reliability. Future work should examine longer deployment horizons, multi-site clinical validation, fairness implications of selective referral, and tighter integration of training-time and inference-time privacy mechanisms.

References

In Neng Chan, Tang Wong, Pak Kin Wong, Tao Yan, In Weng Chan, Hao Ren, and Chon In Chan. Multi-class gastroesophageal reflux disease classification system using deep learning techniques.

In *Proceedings of the 2023 10th International Conference on Biomedical and Bioinformatics Engineering*, pp. 223–229, 2023.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Hailong Ge, Xin Zhou, Yu Wang, Jian Xu, Feng Mo, Chen Chao, Jinzhou Zhu, and Weixin Yu. Development and validation of deep learning models for the multiclassification of reflux esophagitis based on the los angeles classification. *Journal of Healthcare Engineering*, 2023(1):7023731, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 16937–16947, 2020.

Haimei Gong, Liangjun Jiang, Xiaoyang Liu, Yuanqi Wang, Omary Gastro, Lei Wang, Ke Zhang, and Zhen Guo. Gradient leakage attacks in federated learning. *Artificial Intelligence Review*, 56: 1337–1374, 2023.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognition*, 151:110424, 2024.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7436–7456, 2021.

N. Hernandez-Cruz et al. Review of federated learning and machine learning-based methods for medical image analysis. *Big Data and Cognitive Computing*, 8(9):99, 2024.

Georgios Kaissis et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3:473–484, 2021.

Muah Kim, Onur Günlü, and Rafael F. Schaefer. Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication. *arXiv preprint arXiv:2102.04737*, 2021.

Aiden Ko, Aaron Kline, Kaitlyn Dunlap, SaiMourya Surabhi, Parnian Azizian, Peter Y. Washington, and Dennis P. Wall. Abstention and threshold identification for uncertainty management in clinical decision tools: A case study using human-in-the-loop pediatric autism classifiers. In *Pacific Symposium on Biocomputing*, volume 31, pp. 114–129, 2026.

Stefanos Laskaridis, Alexandros Kouris, and Nicholas D. Lane. Adaptive inference through early-exit networks: Design, challenges, and directions. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*, pp. 1–6, 2021.

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.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 2024.

Xiaoyao Ma et al. Heterogeneous federated learning: State-of-the-art and research challenges. *arXiv preprint arXiv:2307.10616*, 2023.

Sarthak Kumar Maharana et al. Continual test-time adaptation in computer vision: Methods, benchmarks, and future directions. *arXiv preprint arXiv:2607.08164*, 2025.

Yoshitomo Matsubara, Marco Levorato, and Francesco Restuccia. Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys*, 55(5): 1–30, 2022.

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 (AISTATS)*, pp. 1273–1282, 2017.

Marziyeh Mohammadi, Mohsen Vejdanihemmat, Mahshad Lotfinia, Mirabela Rusu, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. *npj Digital Medicine*, 9:93, 2026.

Dominic Ng et al. Federated learning for medical imaging radiology. *British Journal of Radiology*, 96(1150):20220890, 2023.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Mélanie Roschewitz et al. Automatic dataset shift identification to support root cause analysis of ai performance drift. *arXiv preprint arXiv:2411.07940*, 2024.

Holli Sargeant, Mackenzie Jorgensen, Arina Shah, Sam Goring, Adrian Weller, and Umang Bhatt. Unequal uncertainty: Rethinking algorithmic interventions for mitigating discrimination from ai. *arXiv preprint arXiv:2508.07872*, 2025.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexander Shen, Luke Francisco, Srijan Sen, and Ambuj Tewari. Exploring the privacy-utility tradeoff in differentially private federated learning for mobile health: A novel approach using simulated privacy attacks. *medRxiv*, 2022.

Vallijah Subasri et al. Diagnosing and remediating harmful data shifts for the responsible deployment of clinical ai models. *JAMA Network Open*, 8(6), 2025.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–362, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Chi-Chih Wang et al. A deep learning model for classification of endoscopic gastroesophageal reflux disease. *International Journal of Environmental Research and Public Health*, 18(5):2428, 2021a.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021b.

Pengju Wang, Bochao Liu, Weijia Guo, Yong Li, and Shiming Ge. Towards personalized federated learning via comprehensive knowledge distillation. *arXiv preprint arXiv:2411.03569*, 2024.

Yanan Wu et al. Test-time domain adaptation by learning domain-aware batch normalization. *arXiv preprint arXiv:2312.10165*, 2023.

Fan Yang et al. Test-time batch normalization. *arXiv preprint arXiv:2205.10210*, 2022.

Hsu-Heng Yen, Hui-Yu Tsai, Chi-Chih Wang, Ming-Chang Tsai, and Ming-Hseng Tseng. An improved endoscopic automatic classification model for gastroesophageal reflux disease using deep learning integrated machine learning. *Diagnostics*, 12(11):2827, 2022.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Junyi Zhu and Matthew B. Blaschko. R-gap: Recursive gradient attack on privacy. *arXiv preprint arXiv:2010.07733*, 2021.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14747–14756, 2019.

A Detailed FedGIM Formulation and Optimization

A.1 Class-Balanced Local Objective

For client k, let $n_{k,c}$ denote the number of local training samples belonging to class c. To reduce the influence of local class imbalance, we use effective-number weighting:

$$w_{k,c} = \frac{1-\beta}{1-\beta^{n_{k,c}}}, \qquad \beta \in [0,1). \tag{15}$$

The weights are normalized over the C classes:

$$\tilde{w}_{k,c} = \frac{C w_{k,c}}{\sum_{j=1}^{C} w_{k,j}}. \tag{16}$$

The resulting weighted local objective is

$$\mathcal{L}_k(\theta, \varphi_k) = -\frac{1}{n_k} \sum_{(x_i,y_i)\in\mathcal{D}_k} \tilde{w}_{k,y_i} \log f_{\theta,\varphi_k}(x_i)_{y_i}. \tag{17}$$

A.2 Proximal Personalized Optimization

FedGIM adds a proximal constraint to the shared parameters:

$$\mathcal{J}_k = \mathcal{L}_k(\theta, \varphi_k) + \frac{\mu}{2}\|\theta - \theta^{(t)}\|_2^2. \tag{18}$$

Each client performs E local optimization epochs beginning from $(\theta^{(t)}, \varphi_k^{(t)})$. The shared parameters are then aggregated according to client sample size:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)}. \tag{19}$$

The personalized parameters remain local:

$$\varphi_k^{(t+1)} = \varphi_k^{(t,E)}. \tag{20}$$

No personalized head parameters are communicated or averaged.

A.3 Relationship to Standard Federated Objectives

FedGIM contains several familiar configurations as special cases:

$$\mu = 0, \quad \varphi_k = \varnothing \Rightarrow \text{FedAvg}, \tag{21}$$

$$\mu > 0, \quad \varphi_k = \varnothing \Rightarrow \text{FedProx}, \tag{22}$$

$$\mu = 0, \quad \varphi_k \neq \varnothing \Rightarrow \text{FedPer-style personalization}. \tag{23}$$

This formulation allows the contribution of the combined training recipe to be separated from the contribution of the inference-management layer.

B LEAKAGE-AWARE PERSONALIZED FL EVALUATION

B.1 PROBLEM DEFINITION

A common personalized FL evaluation procedure evaluates each client's personalized model on that same client's held-out shard:

$$(x, y) \in \mathcal{D}_k^{\text{test}} \quad \Rightarrow \quad f_{\theta,\varphi_k}(x). \tag{24}$$

Under non-IID class distributions, this procedure implicitly supplies the model with the test-client identity through the choice of personalized head. If client k has a characteristic class prior $\mathcal{P}_k(y)$, a personalized predictor can therefore exploit this prior even when image evidence is weak.

We distinguish this paired protocol from pooled evaluation, in which the held-out samples are combined and the true client identity is unavailable during inference.

B.2 LABEL-FREE TEST-TIME ROUTING

For pooled evaluation, the personalized head is selected using the feature-space router introduced in Eq. equation 5. The client centroid is computed exclusively from training data:

$$\mu_k = \frac{\sum_{x_i \in \mathcal{D}_k^{\text{train}}} \phi_{\text{enc}}(x_i)}{\left\| \sum_{x_i \in \mathcal{D}_k^{\text{train}}} \phi_{\text{enc}}(x_i) \right\|_2}. \tag{25}$$

No test label, test-client identifier, or test-label distribution is used by the router.

B.3 SYNTHETIC PRIOR-ONLY CONTROL

To isolate the evaluation mechanism from the visual model, we construct a prior-only predictor that receives no image representation and predicts solely from the empirical client-level class distribution.

Under paired evaluation, the predictor receives the true client assignment through the evaluation protocol. Under pooled evaluation, that information is unavailable.

The experiment therefore tests whether the evaluation protocol itself can produce apparent discrimination without image information. The complete seed-level results, Dirichlet concentration settings, and implementation details are reported in the corresponding experimental results.

C DETAILED INFERENCE MANAGEMENT IMPLEMENTATION

FedIM is implemented as a coordinated four-stage inference pipeline consisting of Monitor, Adapt, Act, and Refer. Unlike a conventional classifier that maps an input directly to a prediction, FedIM maintains an explicit inference state and uses this state to determine whether adaptation, computational escalation, privacy transformation, or human referral is required.

Given an input x, the frozen encoder produces

$$z = \phi_{\text{enc}}(x). \tag{26}$$

The leakage-aware router then selects the appropriate personalized head $f_{\theta,\varphi_{r(x)}}$. The complete inference process is

$$x \to \phi_{\text{enc}}(x) \to \text{Monitor} \to \text{Adapt} \to \text{Act} \to \text{Refer}. \tag{27}$$

The stages are coordinated rather than independent. In particular, Monitor provides a drift signal that can activate Adapt, while Act produces the prediction state used by Refer.

C.1 MONITOR

The Monitor stage evaluates both predictive reliability and distributional change.

Prediction confidence. For a posterior vector $p(x)$, top-1 confidence is

$$c(x) = \max_c p_c(x), \tag{28}$$

while predictive entropy is

$$\mathcal{H}(p) = -\sum_{c=1}^{C} p_c \log p_c. \tag{29}$$

The top-1 margin is

$$m(p) = p_{(1)} - p_{(2)}, \tag{30}$$

where $p_{(1)}$ and $p_{(2)}$ denote the largest and second-largest class probabilities.

Maximum mean discrepancy. Let $\mathcal{R}$ denote the reference feature window and $\mathcal{B}$ the current feature batch. We estimate the unbiased squared MMD as

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{m(m-1)} \sum_{i \neq j} k(r_i, r_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(b_i, b_j) - \frac{2}{mn} \sum_{i,j} k(r_i, b_j). \tag{31}$$

The RBF kernel is

$$k(u, v) = \exp\left(-\gamma \|u - v\|_2^2\right). \tag{32}$$

Kolmogorov–Smirnov monitoring. A two-sample KS test is applied independently to each feature dimension. The significance level is adjusted using Bonferroni correction:

$$\alpha_{\mathrm{KS}} = \frac{\alpha}{d}. \tag{33}$$

The fraction of dimensions exhibiting significant distributional change is

$$q_{\mathrm{KS}} = \frac{1}{d} \sum_{j=1}^{d} \mathbb{1}\left[p_j^{\mathrm{KS}} < \frac{\alpha}{d}\right]. \tag{34}$$

Drift decision. The Monitor stage produces the binary drift signal

$$D = \mathbb{1}\left[\widehat{\mathrm{MMD}}^2 > \tau_{\mathrm{MMD}}\right] \vee \mathbb{1}\left[q_{\mathrm{KS}} > 0.1\right]. \tag{35}$$

Thus, MMD captures aggregate distributional displacement while KS provides dimension-wise evidence of feature-level change.

C.2 ADAPT

The Adapt stage provides label-free test-time adaptation of the routed prediction path. Following TENT (Wang et al., 2021b), adaptation minimizes predictive entropy on the current unlabeled batch.

The visual encoders remain frozen. Adaptation is restricted to the affine parameters of BatchNorm:

$$\Gamma_{\mathrm{adapt}} = \{\gamma, \beta\}. \tag{36}$$

For an unlabeled batch $\mathcal{B}$, predictions are

$$p_i = \mathrm{softmax}\left(f_{\theta, \varphi_k}(z_i)\right). \tag{37}$$

The TENT objective is

$$\mathcal{L}_{\mathrm{TENT}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{c=1}^{C} p_{ic} \log p_{ic}. \tag{38}$$

Only the BatchNorm affine parameters are updated:

$$(\gamma, \beta) \leftarrow (\gamma, \beta) - \eta \nabla_{\gamma, \beta} \mathcal{L}_{\mathrm{TENT}}. \tag{39}$$

The adaptation gate combines the configured TTA state with the Monitor output:

$$a = \mathbb{1}[\text{TTA enabled}] \vee D. \tag{40}$$

Consequently, detected distributional shift can force adaptation even when continuous TTA is disabled.

Multi-view inference. Following adaptation, predictions may be evaluated over V augmented views:

$$\bar{p}(x) = \frac{1}{V} \sum_{v=1}^{V} p(y \mid x^{(v)}). \tag{41}$$

TENT and multi-view inference serve distinct roles: TENT updates the inference parameters, whereas multi-view aggregation stabilizes the final prediction across transformed observations.

C.3 ACT

The Act stage determines the computational and privacy operating point for each case.

Confidence-gated escalation. Each input is first evaluated using the inexpensive quantized head $Q(f_{\theta,\varphi_k})$. Let $\bar{p}^{(\mathrm{cheap})}(x)$ denote its output. Escalation is triggered when its maximum class probability falls below the confidence threshold:

$$e(x) = \mathbb{1}\left[\max_c \bar{p}_c^{(\mathrm{cheap})}(x) < \tau_{\mathrm{conf}}\right]. \tag{42}$$

The selected inference path is then

$$p(x) = (1 - e(x))\bar{p}^{(\mathrm{cheap})}(x) + e(x)\bar{p}^{(\mathrm{full})}(x). \tag{43}$$

Thus, high-confidence cases remain on the inexpensive path, whereas uncertain cases are escalated to full-precision inference.

Personalized–global fusion. When a global model is available, its prediction is combined with the personalized prediction using a fixed fusion weight:

$$p_{\mathrm{fused}}(x) = \frac{1}{2}p(x) + \frac{1}{2}p_\theta(y \mid x). \tag{44}$$

No fusion weight is tuned on the test set.

Inference-time privacy. Let the released pre-noise logit representation be

$$z = \log p_{\mathrm{out}}(x), \tag{45}$$

where $p_{\mathrm{out}} = p_{\mathrm{fused}}$ when a global model is available and $p_{\mathrm{out}} = p$ otherwise.

The released score is perturbed using the Gaussian mechanism:

$$\tilde{z} = z + \mathcal{N}(0, \sigma^2 I), \tag{46}$$

where

$$\sigma = \frac{\Delta\sqrt{2\ln(1.25/\delta)}}{\epsilon}. \tag{47}$$

The privacy operation therefore occurs after inference and does not alter the federated optimization procedure.

C.4 REFER

The Refer stage converts the final prediction into an operational accept-or-refer decision.

The released posterior is

$$\tilde{p} = \mathrm{softmax}(\tilde{z}). \tag{48}$$

Normalized predictive entropy is

$$H_{\mathrm{norm}}(\tilde{p}) = \frac{\mathcal{H}(\tilde{p})}{\log C}. \tag{49}$$

The top-1 margin is

$$m(\tilde{p}) = \tilde{p}_{(1)} - \tilde{p}_{(2)}. \tag{50}$$

We combine entropy and margin into an uncertainty score:

$$u(x) = H_{\text{norm}}(\tilde{p}) + [1 - m(\tilde{p})]. \tag{51}$$

For a calibration set containing N samples and target coverage $\rho \in (0, 1]$, let

$$q_\rho = u_{(\lfloor \rho N \rfloor)} \tag{52}$$

denote the corresponding empirical uncertainty quantile.

The final acceptance rule is

$$\text{acc}(x) = \mathbb{1}[u(x) \le q_\rho] \wedge [\mathcal{H}(\tilde{p}) \le \tau_{\mathcal{H}} \vee m(\tilde{p}) \ge \tau_m]\,. \tag{53}$$

Samples that fail this criterion are referred for human review. The resulting mechanism therefore converts uncertainty into an explicit selective-prediction policy rather than forcing every case through an automated decision.

D Privacy Mechanism and Output-Level Guarantee

The inference-time privacy mechanism operates on the released score $\tilde{z}$ defined in Eq. equation 46. For neighboring inputs x, x' satisfying

$$\|z(x) - z(x')\|_2 \le \Delta, \tag{54}$$

the Gaussian mechanism provides the output-level guarantee

$$\Pr[\tilde{z}(x) \in S] \le e^\epsilon \Pr[\tilde{z}(x') \in S] + \delta, \qquad \forall S. \tag{55}$$

This guarantee applies to the released inference output and should not be interpreted as a training-time privacy guarantee for the learned parameters. The federated training procedure itself is not claimed to satisfy differential privacy unless explicitly stated elsewhere in the experimental protocol.

E Selective Prediction and Coverage Analysis

The Refer stage produces a binary acceptance decision for each test sample. For a test set of N samples, empirical coverage is

$$\text{Coverage} = \frac{1}{N} \sum_{i=1}^{N} \text{acc}(x_i). \tag{56}$$

The corresponding selective risk is computed over accepted predictions:

$$\text{Risk} = \frac{\sum_{i=1}^{N} \mathbb{1}[\text{acc}(x_i)] \mathbb{1}[\hat{y}_i \ne y_i]}{\sum_{i=1}^{N} \mathbb{1}[\text{acc}(x_i)]}. \tag{57}$$

Coverage–risk curves are obtained by varying the target coverage parameter ρ. This evaluates whether the Refer stage can reduce the error rate among automatically accepted cases by referring uncertain samples.

F Computational and Deployment Analysis

The computational behavior of FedIM depends on the proportion of inputs that are escalated from the quantized path to the full-precision path.

Table 3: Inference Management configuration used in the experiments.

Component	Configuration
Feature representation	Frozen encoder representation
Personalized model	Client-specific prediction head
Drift detection	MMD + per-dimension KS
MMD kernel	RBF
Multiple testing	Bonferroni correction
TTA method	TENT-style entropy minimization
Adapted parameters	BatchNorm γ, β only
TTA optimizer	Adam
TTA learning rate	10^{-3}
TTA steps	3
Multi-view inference	V augmented views
Cheap inference	INT8 prediction head
Escalation	Confidence gated
Global/personalized fusion	Fixed $0.5/0.5$
Privacy mechanism	Gaussian mechanism
Selective prediction	Entropy + top-1 margin
Referral	Coverage and uncertainty gated

The escalation rate is

$$\bar{e} = \frac{1}{N} \sum_{i=1}^{N} e(x_i). \tag{58}$$

For an individual sample, the inference latency is modeled as

$$\ell(x) = \ell_{\text{enc}} + \ell_{\text{cheap}} + e(x)\ell_{\text{full}} + \ell_{\text{drift}}. \tag{59}$$

The corresponding average throughput is therefore characterized by

$$\text{Throughput} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} \ell(x_i)}. \tag{60}$$

This analysis reports the realized mixed-path operating cost rather than assuming that every sample requires the full inference path.

G Implementation Configuration

The principal implementation settings used throughout the Inference Management pipeline are summarized in Table 3. These settings define the operational configuration of the Monitor, Adapt, Act, and Refer stages, including drift detection, test-time adaptation, computational escalation, privacy transformation, and selective prediction. Unless otherwise specified in an ablation, the same configuration is used across compared models to ensure that differences in performance are attributable to the evaluated method rather than to stage-specific tuning.

H Additional Experiments

H.1 Privacy–Utility Trade-off and Response to Feature-Space Covariate Drift

We evaluate two complementary aspects of inference-time management: the effect of logit perturbation under different privacy budgets, and the response of the monitoring mechanism to controlled feature-space covariate drift.

As the privacy budget increases, predictive utility progressively approaches the non-private configuration. At $\epsilon = 0.5$, AUROC is 0.5537 ± 0.0207 and F1 is 0.2148 ± 0.0268. At $\epsilon = 8$, these

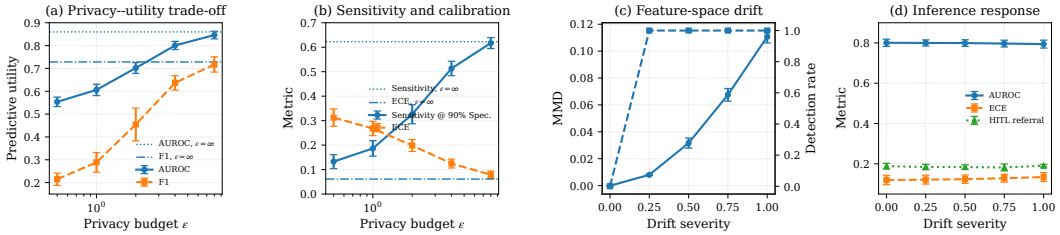


Figure 6: Privacy–utility behavior under inference-time logit perturbation (left) and inference-management response to controlled feature-space covariate drift (right). Increasing the privacy budget progressively restores predictive and calibration performance, while increasing drift severity is reflected in rising MMD, detection behavior, and calibration metrics.

values increase to 0.8462 ± 0.0168 and 0.7175 ± 0.0335, respectively. The non-private configuration obtains AUROC 0.8599 ± 0.0020 and F1 0.7284 ± 0.0087. ECE follows the opposite trend, decreasing from 0.3124 ± 0.0356 at $\epsilon = 0.5$ to 0.0789 ± 0.0146 at $\epsilon = 8$ and 0.0612 ± 0.0118 without differential privacy.

Under controlled feature-space covariate drift, the measured MMD increases with drift severity, from -0.0001 ± 0.0004 at severity 0.00 to 0.1107 ± 0.0046 at severity 1.00. The drift detector identifies all five evaluated seeds from severity 0.25 onward, while no drift is detected at severity 0.00. AUROC remains close to 0.80 over the controlled drift range, while ECE increases as the distribution becomes more shifted.

H.2 Evaluation Leakage in Personalized Federated Learning

We examine the effect of evaluation protocol on personalized federated models. The comparison distinguishes pooled evaluation from client-matched evaluation, where the latter provides the model with the identity of the originating client through the choice of personalized head. For FedGIM, AUROC changes from 0.8512 ± 0.0124 under pooled evaluation to 0.9387 ± 0.0215 under paired client-matched evaluation, corresponding to an AUROC difference of 0.0875 ± 0.0312. Sensitivity at 90% specificity changes from 0.6087 ± 0.0318 to 0.8412 ± 0.0483.

A similar effect is observed for personalized baselines. FedPer changes from 0.7793 ± 0.0299 to 0.9346 ± 0.0295 AUROC, corresponding to an observed difference of 0.1552 ± 0.0570. These results motivate the pooled, label-free routing protocol used throughout the main evaluation.

H.3 Prior-Only Sanity Check

To isolate the contribution of client-level label priors, we evaluate a predictor that uses no image information. Under pooled evaluation, the prior-only predictor remains close to chance with AUROC 0.5057 ± 0.0219 across the reported Dirichlet settings. In contrast, the paired evaluation protocol produces substantially higher AUROC values, including 0.8717 ± 0.0852 at Dirichlet concentration 0.05 and 0.7563 ± 0.0670 at concentration 0.50.

Figure 7 visualizes this discrepancy. The result provides a direct sanity check that client-level label-prior information can create apparent predictive performance when evaluation exposes client identity, even in the absence of image information.

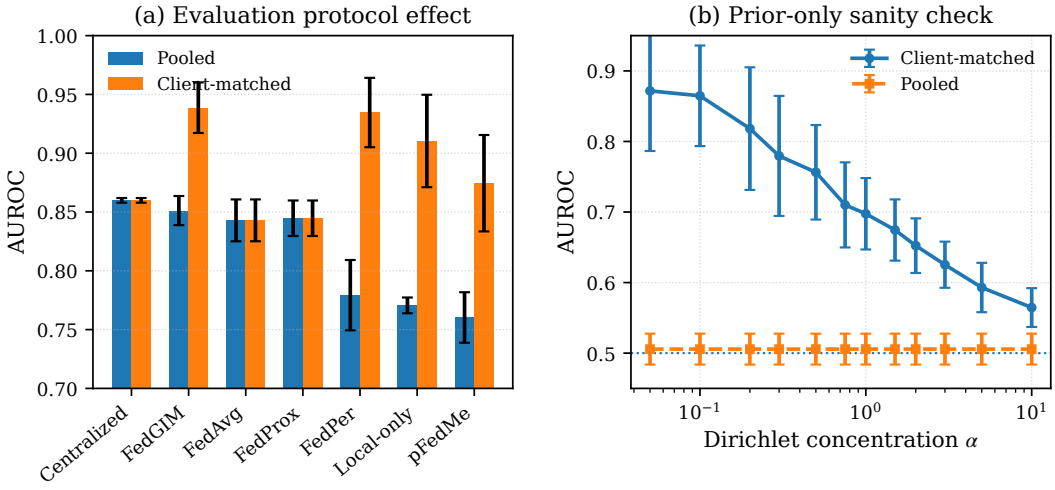


Figure 7: Effect of evaluation protocol on personalized federated learning. Client-matched evaluation is compared with pooled evaluation, with a prior-only predictor providing a label-free sanity check.