Bayes-pFCL: Bayesian Personalized Federated Continual Learning

Qingyang Yu
Department of Electrical and Computer Engineering
Stevens Institute of Technology
qyu13@stevens.edu

Yang Hua
School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast
Y.Hua@qub.ac.uk

Hao Wang
School of Information Sciences
University of Illinois Urbana-Champaign
hwangml@illinois.edu

Yue Ning
School of Computing
Stevens Institute of Technology
yue.ning@stevens.edu

Qizhen Zhang
Department of Computer Science
University of Toronto
qz@cs.toronto.edu

Hao Wang
Department of Electrical and Computer Engineering
Stevens Institute of Technology
hwang9@stevens.edu

Abstract

Personalized federated continual learning (pFCL) alleviates catastrophic forgetting across tasks and data heterogeneity across clients. We view both challenges as interference between knowledge and propose Bayesian frameworks that define knowledge as posterior belief and quantify two types of interference: intra-model interference, measuring task-induced posterior drift, and inter-model interference, measuring global–personalized posterior dependence. These interference measures bound expected forgetting loss and the expected excess positive loss gap over a decoupled-posterior baseline, respectively. We then develop an interference-regularized local objective to guide personalization under catastrophic forgetting and data heterogeneity. The framework unifies standard FCL and pFL categories. Experiments on synthetic and real-world benchmarks demonstrate improved model performance over state-of-the-art pFCL methods.

1 Introduction

Federated continual learning (FCL) has emerged as an effective paradigm for privacy-preserving learning in healthcare [63, 63, 21, 48] and finance [5, 62, 45, 5], where clients learn a general model from streaming data without sharing raw data. In FCL, each client's streaming data (represented as the task sequence) carries *knowledge*, while the global model integrates such knowledge across clients and over time. However, in real-world applications, knowledge often exhibits high heterogeneity between clients and between tasks along each client's task sequence [22, 29, 10, 56, 28]. As a result, knowledge both learned from other clients and the current task impairs a client's retention of knowledge from previous tasks, as illustrated in Figure 1. Thus, integrating heterogeneous knowledge across clients and tasks requires personalization along each client's task sequence, which is beyond the capability scope of the single-global-model paradigm of FCL and thereby yields suboptimal client-level performance. To address this limitation, personalized federated continual learning (pFCL) extends FCL by allowing each client to personalize its local model for task sequences while retaining the privacy-preserving and collaborative continual-learning benefits of FCL.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

pFCL primarily focuses on personalization under two challenges, *catastrophic forgetting* and *data heterogeneity*, which respectively arise from intra-model and inter-model *interference* between knowledge: (1) catastrophic forgetting reflects interference within each client's task sequence: due to no access to previous tasks in continual learning (CL)-style scenarios, knowledge learned from the current task overwrites knowledge learned from previous tasks during optimization, which we call *intra-model interference*; and (2) data heterogeneity reflects interference across clients' task sequences: knowledge from heterogeneous clients is integrated during global aggregation to serve all clients, rather than each client's personalization needs, which we call *inter-model interference*. A rich literature of pFCL has been proposed by exploring several orthogonal methodologies, *e.g.*, regularization-based methods [13, 46, 55, 57, 59], and replay-based methods [62]. Existing studies mainly alleviate these challenges as follows: approximate knowledge with surrogates, *e.g.*, prompts [57] or classifiers [31], then decouple invariant and specific information from these surrogates and use such information for regularization or replay. However, these approaches often rely on heuristic knowledge proxies, lacking a principled mathematical formulation of *knowledge* and largely overlooking *interference*. Such a gap prevents prior studies from explicitly characterizing how knowledge interferes at intra- and inter-model level, leading to suboptimal model performance. Thus, a natural question arises: *can we develop an intuitive yet mathematical framework for knowledge and interference in pFCL?*

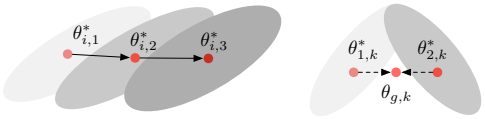


Figure 1: Shaded ellipses denote task-specific posterior concentrations, and red points denote task-wise optima. **Left:** Task shifts move optimum $\theta^*_{i,3}$ away from prior tasks, impairing knowledge retention on tasks 1–2. **Right:** Aggregation pulls personalized optima $\theta^*_{i,k}$ toward $\theta_{g,k}$, weakening knowledge retention.

The above question leads us to revisit the domain where *knowledge* and *interference* were first formalized, human cognition (HC) [4, 47, 2, 42]. In HC, knowledge is the probability of retrieving a memory chunk under a cue, *e.g.*, retrieving "fruit" from the cue "apple," whereas interference is the uncertainty caused by multiple chunks sharing the same cue, *e.g.*, associating "apple" with both "fruit" and "corporation." Inspired by such a probabilistic view, and building upon the general principles of Bayesian deep learning (BDL) [52, 53], which integrates deep neural networks with probabilistic modeling and inference, we develop a static Bayesian framework to formally define and quantify knowledge and interference. Then, to address the inability of the static Bayesian framework to capture temporal evolution during optimization, we extend it to a dynamic version that characterizes how knowledge and interference evolve throughout training. Theoretically, intra-model interference bounds expected forgetting loss, while population inter-model interference bounds the expected excess positive personalized–global loss gap over a decoupled-posterior baseline. Our framework also naturally generalizes to several personalized federated learning (pFL) categories and to FCL. We then use the proposed intra- and inter-model interference to regularize catastrophic forgetting and data heterogeneity, respectively, thereby guiding personalization in pFCL.

Our contributions are as follows:

- Inspired by framing knowledge and interference as probabilistic models in HC, we provide the first mathematical and principled framework of "knowledge" and "interference" in pFCL, and develop a general method, dubbed Bayes-pFCL, for handling personalization under catastrophic forgetting and data heterogeneity in pFCL.
- We derive interference-based bounds on expected forgetting loss and on the expected excess positive personalized–global loss gap over a decoupled-posterior baseline.
- Experiments on synthetic and real-world datasets, *e.g.*, *TCGA-BRCA*, show that Bayes-pFCL effectively improves performance over state-of-the-art (SOTA) pFCL.

2 Related Work

Federated Continual Learning (FCL) aims to mitigate *catastrophic forgetting* from sequential task arrival and *statistical heterogeneity* across clients. Existing FCL methods mainly include: (1) regularization-based methods [11, 20, 32, 50]; (2) task-based methods [56]; (3) replay-based methods [1, 44, 41]; and (4) other related approaches [7, 18, 17]. Regularization-based methods penalize disruptive updates to preserve past knowledge; task-based methods exploit task structures to mitigate

forgetting; and replay-based methods maintain historical knowledge through stored, generated, or reconstructed samples, often yielding strong performance in class-incremental settings [51].

Personalized Federated Continual Learning (pFCL) extends standard FCL by requiring learned models to preserve client-specific knowledge while benefiting from cross-client collaboration. The training objective must therefore maintain performance on previously seen tasks and improve personalized generalization for each client. Achieving this objective requires balancing the stability–plasticity trade-off: each client must remain plastic enough to acquire knowledge from new tasks and stable enough to preserve previously learned knowledge. Existing pFCL studies can be broadly categorized into two groups: (1) regularization-based methods, including pFedDIL [31], FedMGP [57], DCFCL [55], FedAGC [59], and Adp-PFCL [13], constrain model updates or transfer knowledge across tasks and clients through regularization, prompting, distillation, or gradient correction; (2) replay-based methods, such as Meta-RBCIL [62], mitigate forgetting by synthesizing historical information during sequential personalized training. Beyond federated learning (FL), Bayesian domain adaptation methods such as VDI [54] model domain-specific characteristics probabilistically to improve adaptation and interpretability. These methods demonstrate the potential of personalization in federated continual settings, but they also expose the unique difficulty of mitigating knowledge interference in pFCL due to the lack of mathematically quantifying knowledge and interference. We address such a gap by explicitly formulating intra- and inter-model interference, enabling personalization to mitigate catastrophic forgetting and data heterogeneity more effectively.

Human Cognition (HC) has long studied how the human mind accumulates knowledge while managing interference among competing associations. In *adaptive control of thought—rational (ACT-R)* [4] and related memory theories[3], knowledge is defined through the probability of retrieving target information given a cue, while interference arises from the *fan effect* [2]: one cue linked to multiple memories activates competing associations and reduces target retrieval probability. For example, when the cue "apple" links to both a fruit and a technology company, target retrieval becomes more difficult because competing associations are simultaneously activated. This problem has a natural parallel to knowledge and interference in machine learning (ML) [36, 19]. Beyond *ACT-R*, global matching models such as *search of associative memory (SAM)* [42] and *retrieving effectively from memory (REM)* [47] provide similar probabilistic accounts of retrieval and interference through search- and Bayesian likelihood mechanisms. These cognitive theories ground our framework by motivating a probabilistic view of knowledge and interference.

3 Method

Problem Setup. Consider N clients $\{C_i\}_{i\in[N]}$, where each client i observes a sequence of K tasks $\{T_{i,k}\}_{k\in[K]}$ sequentially. The dataset of task k on client i is denoted by $\mathcal{D}_{i,k} = \{(x, y) \mid (x, y) \sim T_{i,k}\}$, where $T_{i,k}$ is the underlying data distribution of k-th task of i-th client.

Let θ_g denote the global model and θ_i denote the personalized model of client i. In pFCL, the goal is to learn personalized models $\{\theta_i\}_{i=1}^N$, induced by θ_g, that achieve the best average performance over all tasks observed by each client. The corresponding empirical objective of client i remains

$$\min_{\{\theta_i\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}), \tag{1}$$

subject to sequential task arrival and federated communication constraints. Eqn. 1 represents the general empirical objective of pFCL, inherited from the empirical objective of pFL, where the final performance is measured on the personalized models $\{\theta_i\}_{i=1}^N$. For specific pFL categories, Eqn. 1 may include additional category-dependent terms, *e.g.*, a regularization term $\mathcal{R}(\theta_g, \theta_i)$ in regularization-based pFL [30]. We defer the discussion of the compatibility of our Bayesian framework with different pFL categories and FCL in Appendix D.

3.1 Knowledge and Interference

Inspired by the cognitive architecture ACT-R, which models memory through probabilistic associations, we formally define the general concept of "knowledge" as follows:

Definition 3.1 (**Knowledge**). *Given a model parameterized by θ and a dataset $\mathcal{D}$, the* knowledge *encoded by θ with respect to $\mathcal{D}$ is defined as the likelihood that θ assigns to $\mathcal{D}$:*

$$\mathcal{K}(\theta; \mathcal{D}) := p(\mathcal{D} \mid \theta).$$

Drawing further inspiration from ACT-R, interference in HC can be viewed as an algebraic change in a scalar retrieval quantity, such as an increase or decrease in the probability of retrieving a target memory under a given cue. However, defining interference as a scalar-level operation is not directly compatible with the notion of knowledge in Def. 3.1. The mismatch arises from two perspectives: (1) from the *intra-model perspective*, knowledge is induced by observed *task* and represented as a likelihood belief over high-dimensional model parameters, rather than as a single retrieval probability. Thus, interference within one knowledge-bearing parameter should capture how newly observed task reshapes the knowledge associated with that parameter; and (2) from the *inter-model perspective*, interference also arises from shared information across different knowledge-bearing parameters, *e.g.*, models, rather than solely from changes within one model. Such two perspectives require a probabilistic measure that captures both how new task changes knowledge within a parameter and how different knowledge-bearing parameters depend on each other. Conditional mutual information (MI) provides such a measure since it quantifies how much one random variable reduces uncertainty about another under given observations. Therefore, we use conditional MI to quantify interference in two complementary ways: intra-level interference, which measures how newly observed task changes the knowledge associated with a model parameter, and inter-level interference, which measures how strongly different knowledge-bearing parameters are statistically coupled.

We formally define the general concept of "interference" from two complementary perspectives:

Definition 3.2 (**Interference**). *Given knowledge in Def. 3.1, we quantify* interference *as follows:*

- ***Intra-model interference.*** *Given previously observed task, e.g., $\mathcal{D}_1$, and newly observed task, e.g., $\mathcal{D}_2$, the intra-model interference induced by $\mathcal{D}_2$ on $p(\mathcal{D}_1 \mid \theta)$ is defined as the conditional information gain about the knowledge-bearing parameter θ:*

$$\mathcal{I}^{\text{intra}}_{\theta; \mathcal{D}_2 \to \mathcal{D}_1} := I(\theta; \mathcal{D}_2 \mid \mathcal{D}_1) = \mathbb{E}_{p(\mathcal{D}_1, \mathcal{D}_2)} \left[\mathrm{KL} \left(p(\theta \mid \mathcal{D}_1, \mathcal{D}_2) \,\|\, p(\theta \mid \mathcal{D}_1) \right) \right]. \tag{2}$$

- ***Inter-model interference.*** *Given two knowledge-bearing parameters θ_1 and θ_2 and task data $\mathcal{D}$, the population inter-model interference between θ_2 and θ_1 is defined as:*

$$\mathcal{I}^{\text{inter}}_{\theta_1 \leftrightarrow \theta_2; \mathcal{D}} := I(\theta_1; \theta_2 \mid \mathcal{D}) = \mathbb{E}_{p(\mathcal{D})} \left[\mathrm{KL} \left(p(\theta_1, \theta_2 \mid \mathcal{D}) \,\|\, p(\theta_1 \mid \mathcal{D}) p(\theta_2 \mid \mathcal{D}) \right) \right]. \tag{3}$$

We next use the Bayesian framework to interpret *knowledge* (Def. 3.1) and *interference* (Def. 3.2) specialized in pFCL scenarios from both static and dynamic views.

Static Bayesian Views of *Knowledge* and *Interference* in pFCL. pFCL scenarios involve multiple variables (*e.g.*, θ_g, $\{\theta_i\}_{i=1}^N$) with complex couplings between θ_g and $\{\theta_i\}_{i=1}^N$. Such couplings make an isolated view of each θ_i insufficient for interpreting knowledge and interference. Bayesian deep learning (BDL) [52, 53] provides a general framework for integrating probabilistic graphical models with deep neural networks, allowing model parameters and their dependencies to be represented probabilistically. We therefore adopt a hierarchical Bayesian view [14], moving from the *whole* joint knowledge of $\{\theta_g, \{\theta_i\}_{i=1}^N\}$, to the *coupling*, *i.e.*, $p(\theta_i \mid \theta_g)$. Under this view, we specialize Def. 3.1 from its general probabilistic form to the pFCL setting. In pFCL, the specialized definition must respect two structural constraints: sequential task arrival (*i.e.*, CL) and personalized federated model structure (*i.e.*, pFL).

First, the CL setup changes the knowledge object from a single-task likelihood to accumulated task evidence. For client-level CL, knowledge after task K is associated with the sequential evidence $p(\mathcal{D}_{1:K} \mid \theta)$ rather than a single-task likelihood $p(\mathcal{D} \mid \theta)$, where $\mathcal{D}_{1:K} = \{\mathcal{D}_k\}_{k=1}^K$. Because the learner observes tasks sequentially and usually cannot revisit $\mathcal{D}_{1:K-1}$ when training on task K, such accumulated likelihood is not directly computable during optimization. With a fixed prior $p(\theta)$ [33, 6, 15, 24], we therefore use the posterior as a tractable surrogate for accumulated knowledge:

$$p(\theta \mid \mathcal{D}_{1:K}) \propto p(\theta) p(\mathcal{D}_{1:K} \mid \theta). \tag{4}$$

Second, the pFL setup changes the knowledge carrier from one parameter to a coupled collection of global and personalized parameters. The global parameter θ_g represents shared knowledge, while

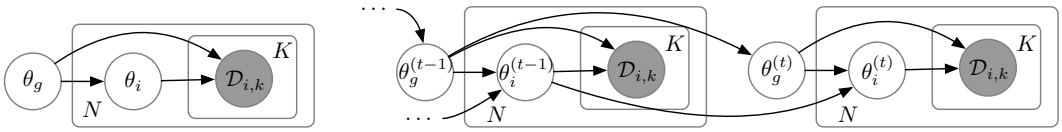


Figure 2: **Left:** Static Bayesian probabilistic graphical model (PGM) of knowledge in pFCL, where θ_g, θ_i encode client i's knowledge over tasks $\{\mathcal{D}_{i,k}\}_{k=1}^K$. **Right:** Dynamic Bayesian PGM of knowledge in pFCL, where $\theta_g^{(t)}, \theta_i^{(t)}$ evolve from round $(t-1)$ after observing $\mathcal{D}_{i,k}$.

each personalized parameter θ_i represents client-specific knowledge coupled through θ_g. Thus, in pFCL, knowledge of $\{\theta_g, \{\theta_i\}_{i=1}^N\}$ on $\mathcal{D}_{1:N,1:K}$ is represented by the joint posterior

$$\mathcal{K}(\theta_g, \{\theta_i\}_{i=1}^N; \mathcal{D}_{1:N,1:K}) := p\big(\theta_g, \{\theta_i\}_{i=1}^N \mid \mathcal{D}_{1:N,1:K}\big)\,. \tag{5}$$

Under the static Bayesian formulation, Bayes' rule factorizes the joint posterior as

$$p(\theta_g, \{\theta_i\}_{i=1}^N \mid \mathcal{D}_{1:N,1:K}) \propto p(\theta_g) \prod_{i=1}^N p(\theta_i \mid \theta_g) \prod_{k=1}^K \prod_{i=1}^N p(\mathcal{D}_{i,k} \mid \theta_g, \theta_i), \tag{6}$$

where $p(\theta_i \mid \theta_g)$ models the coupling between the global knowledge and personalized knowledge, and $p(\mathcal{D}_{i,k} \mid \theta_g, \theta_i)$ models the dependence of client i's task-k data on both components; see Figure 2.

Bayes' rule updates the joint posterior recursively across tasks as

$$p(\theta_g, \{\theta_i\}_{i=1}^N \mid \mathcal{D}_{1:N,1:K}) \propto p(\theta_g, \{\theta_i\}_{i=1}^N \mid \mathcal{D}_{1:N,1:K-1}) \prod_{i=1}^N p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i). \tag{7}$$

The posterior accumulated from previous tasks therefore serves as the prior for incorporating the current task across all clients. Thus, pFCL recursively updates a joint posterior over global and personalized *knowledge* on the current task k.

Via the *Laplace approximation*, we approximate the posterior after task $K-1$ by a Gaussian centered at its MAP estimate:

$$p(\theta_g, \{\theta_i\}_{i=1}^N \mid \mathcal{D}_{1:N,1:K-1}) \approx \mathcal{N}\big(z; \mu_{K-1}, \Lambda_{K-1}^{-1}\big)\,, \tag{8}$$

where

$$z := \begin{bmatrix}\theta_g^\top & \theta_1^\top & \cdots & \theta_N^\top\end{bmatrix}^\top,$$
$$\mu_{K-1} := \begin{bmatrix}\mu_{g,K-1}^\top & \mu_{1,K-1}^\top & \cdots & \mu_{N,K-1}^\top\end{bmatrix}^\top,$$

and the precision matrix follows the same parameter ordering:

$$\Lambda_{K-1} = \begin{bmatrix} \Lambda_{gg,K-1} & \Lambda_{g1,K-1} & \cdots & \Lambda_{gN,K-1} \\ \Lambda_{1g,K-1} & \Lambda_{11,K-1} & \cdots & \Lambda_{1N,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{Ng,K-1} & \Lambda_{N1,K-1} & \cdots & \Lambda_{NN,K-1} \end{bmatrix}.$$

Then, the log posterior at task K is approximately

$$\log p(z \mid \mathcal{D}_{1:N,1:K}) = \sum_{i=1}^N \log p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i) - \frac{1}{2}(z-\mu_{K-1})^\top \Lambda_{K-1}(z-\mu_{K-1}) + C,$$

where C is a constant independent of z. Then, we have the mode μ_K as

$$\begin{aligned}\mu_K &= \arg\max_z \log p(z \mid \mathcal{D}_{1:N,1:K}) \\ &\approx \arg\max_z \left[\log p(\mathcal{D}_{1:N,K} \mid \theta_g, \theta_i) - \frac{1}{2}(z-\mu_{K-1})^\top \Lambda_{K-1}(z-\mu_{K-1})\right]\end{aligned}$$

Theorem 3.3 (**Laplace representation of global–personalized coupling**). *Suppose the negative log-posterior induced by Eqn. 6 is twice continuously differentiable in a neighborhood of its MAP estimate μ_K. Under the Laplace approximation*

$$p(z \mid \mathcal{D}_{1:N,1:K}) \approx \mathcal{N}(z; \mu_K, \Lambda_K^{-1}),$$

the (g,i)-th off-diagonal precision block is given by

$$\Lambda_{gi,K} = \nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(z \mid \mathcal{D}_{1:N,1:K}) \right]_{z=\mu_K}.$$

Then we have

$$\Lambda_{gi,K} = \underbrace{\nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(\theta_i \mid \theta_g) \right]_{z=\mu_K}}_{\textit{prior global–personalized coupling}} + \sum_{k=1}^{K} \underbrace{\nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(\mathcal{D}_{i,k} \mid \theta_g, \theta_i) \right]_{z=\mu_K}}_{\textit{task-induced coupling}}.$$

Therefore, the hierarchical coupling term $p(\theta_i \mid \theta_g)$ is locally represented, under the Laplace approximation, by its cross-curvature contribution to the joint precision block $\Lambda_{gi,K}$.

As a result, the (g,i)-th precision block can be updated online as

$$\Lambda_{gi,K} = \Lambda_{gi,K-1} + H_{gi,K}, \qquad H_{gi,K} = \nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i) \right] \Big|_{z=\mu_K}.$$

By Theorem 3.3, the off-diagonal precision block $\Lambda_{gi,K}$ locally represents the direct Gaussian coupling between the global model θ_g and client i's personalized model θ_i. Def. 3.2 gives a general interference principle, but applying it to the full pFCL posterior would entangle direct coupling with indirect dependencies from other clients and previous tasks. Such an aggregate MI cannot identify which interference source client i should control. Thus, we specialize the general interference principle into two actionable quantities:

- **Intra-model interference**, which measures how the current task changes the accumulated posterior knowledge carried by client i's personalized model θ_i. [1] Let $q_{i,k} := p(\theta_i \mid \mathcal{D}_{1:N,1:k})$, for $k \in \{K-1, K\}$, be the personalized marginals of the joint posterior. At the population level, exact Bayesian updating gives

$$\begin{aligned} \mathcal{I}^{\text{intra}}_{\theta_i;[N],K\to<K} &:= I(\theta_i; \mathcal{D}_{1:N,K} \mid \mathcal{D}_{1:N,1:K-1}) \\ &= \mathbb{E}_{\mathcal{D}_{1:N,1:K}} [\mathrm{KL}(q_{i,K} \| q_{i,K-1})]. \end{aligned} \tag{9}$$

  For observed data, the KL measures the change in client i's personalized marginal posterior. Its Gaussian approximation is a computable surrogate; Appendix F.2 gives its Gaussian closed form, while Appendix A.2 distinguishes realized information gain from population MI.
- **Inter-model interference**, which isolates the coupling between θ_g and θ_i after conditioning on the remaining personalized models. Under the *Laplace* approximation $p(z \mid \mathcal{D}_{1:N,1:K}) \approx \mathcal{N}(z; \mu_K, \Lambda_K^{-1})$, let $\theta_{-i} := \{\theta_j\}_{j \neq i}$ and

$$\Lambda_{\{g,i\},K} = \begin{bmatrix} \Lambda_{gg,K} & \Lambda_{gi,K} \\ \Lambda_{ig,K} & \Lambda_{ii,K} \end{bmatrix}.$$

  The population interference and its Gaussian surrogate are

$$\begin{aligned} \mathcal{I}^{\text{inter}}_{\theta_i \leftrightarrow \theta_g; \mathcal{D}_{1:N,1:K}} &:= I(\theta_i; \theta_g \mid \theta_{-i}, \mathcal{D}_{1:N,1:K}) \\ &\approx \frac{1}{2} \mathbb{E}_{\mathcal{D}_{1:N,1:K}} \left[\log |\Lambda_{gg,K}| + \log |\Lambda_{ii,K}| - \log |\Lambda_{\{g,i\},K}| \right]. \end{aligned} \tag{10}$$

  The precision blocks depend on the observed history. Averaging the realized Gaussian penalty yields the CMI of that Gaussian family under the same data law (Appendix F.2).

Dynamic Bayesian Views of *Knowledge* and *Interference* in pFCL. The static Bayesian framework simply compresses all communication rounds within task K into a task-level posterior, thereby

[1] For notation simplicity, we use $\mathcal{I}^{\text{intra}}_{\theta_i;[N],K\to<K}$ as shorthand for $\mathcal{I}^{\text{intra}}_{\theta_i;\mathcal{D}_{1:N,K}\to\mathcal{D}_{1:N,1:K-1}}$ when necessary.

overlooking aggregation-induced effects. Thus, we extend the static framework into a dynamic one by introducing the communication round $t \in [\mathcal{T}_K]$ as a within-task temporal variable; see Figure 2.

For fixed task K, we suppress its index on model parameters, as in Figure 2, writing $\theta_g^{(t)} := \theta_{g,K}^{(t)}$ and $\theta_i^{(t)} := \theta_{i,K}^{(t)}$. Let $\Theta_K^{(t)} := \big(\theta_g^{(t)}, \theta_1^{(t)}, \ldots, \theta_N^{(t)}\big)$ collect the global and personalized models at round t, and write $\theta_{-i}^{(t)} := \{\theta_j^{(t)}\}_{j \neq i}$ for the remaining personalized models. The dynamic posterior is initialized by the static posterior accumulated before task K: $\pi(\Theta_K^{(0)}) := p(\Theta_K^{(0)} \mid \mathcal{D}_{1:N,1:K-1})$. Then, we have

$$\textbf{Knowledge} : \mathcal{K}\big(\Theta_K^{(t)}; \mathcal{D}_{1:N,1:K}\big) := p\big(\Theta_K^{(t)} \big| \Theta_K^{(0:t-1)}, \mathcal{D}_{1:N,1:K}\big), \tag{11}$$

$$\textbf{Factorization} : \mathcal{K}\big(\Theta_K^{(t)}; \mathcal{D}_{1:N,1:K}\big) \propto \pi\big(\Theta_K^{(0)}\big) \prod_{s=1}^{t} p\big(\Theta_K^{(s)} \big| \Theta_K^{(s-1)}\big) p\big(\mathcal{D}_{1:N,K} \big| \theta_g^{(s)}, \theta_i^{(s)}\big), \tag{12}$$

$$\textbf{Intra-Model Interference} : \mathcal{I}^{\text{intra}}_{\theta_i^{(t)};[N],K\to<K} := I\Big(\theta_i^{(t)}; \mathcal{D}_{1:N,K} \,\Big|\, \mathcal{D}_{1:N,1:K-1}\Big), \tag{13}$$

$$\textbf{Inter-Model Interference} : \mathcal{I}^{\text{inter}}_{\theta_i^{(t)} \leftrightarrow \theta_g^{(t)}; \mathcal{D}_{1:N,1:K}} := I(\theta_i^{(t)}; \theta_g^{(t)} \mid \theta_{-i}^{(t)}, \mathcal{D}_{1:N,1:K}). \tag{14}$$

Both dynamic interference terms are population CMI. Appendix F derives their realized Gaussian surrogates and distinguishes data averaging from trajectory marginalization.

3.2 Interference-Based Upper Bounds for Catastrophic Forgetting and Data Heterogeneity

We next bound forgetting-induced loss and the excess positive personalized–global loss gap over a decoupled-posterior baseline in pFCL. Detailed analysis and proofs are available in Appendix A.

Assumption 3.4 ($L_{i,K}^F$**-Lipschitz).** For client i at task $K \geq 2$, fix $\mathcal{D}_{1:N,1:K-1}$ and define the averaged local old-task loss

$$f_{i,K}(\theta_i) := \frac{1}{K-1} \sum_{k=1}^{K-1} \mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}), \tag{15}$$

and the corresponding positive expected loss gap

$$F_{i,K} := \big(\mathbb{E}_{q_{i,K}}[f_{i,K}(\theta_i)] - \mathbb{E}_{q_{i,K-1}}[f_{i,K}(\theta_i)]\big)_+ . \tag{16}$$

Assume that $f_{i,K}$ is $L_{i,K}^F$-Lipschitz with respect to θ_i, i.e.,

$$|f_{i,K}(\theta_i) - f_{i,K}(\theta_i')| \leq L_{i,K}^F \, \|\theta_i - \theta_i'\| , \quad \forall \theta_i, \theta_i'. \tag{17}$$

Assumption 3.5 ($L_{i,K}^H$**-Lipschitz).** For client i at task $K \geq 2$, write $\mathcal{D} := \mathcal{D}_{1:N,1:K}$. For each observed history, define the averaged personalized–global old-task loss gap

$$h_{i,K}(\theta_g, \theta_i) := \frac{1}{K-1} \sum_{k=1}^{K-1} \left[\mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}) - \mathcal{L}_i(\theta_g; \mathcal{D}_{i,k})\right]. \tag{18}$$

For fixed θ_{-i}^*, let

$$Q_{i,K}^{\theta_{-i}^*} := p\big(\theta_g, \theta_i \,\big|\, \theta_{-i}^*, \mathcal{D}_{1:N,1:K}\big) , \tag{19}$$

and define the posterior-expected positive loss gap

$$H_{i,K} := \mathbb{E}_{\theta_{-i} | \mathcal{D}_{1:N,1:K}} \left[\mathbb{E}_{Q_{i,K}^{\theta_{-i}}} \Big[\big(h_{i,K}(\theta_g, \theta_i)\big)_+ \Big] \right]. \tag{20}$$

The data dependence of $h_{i,K}$ and $H_{i,K}$ is suppressed. Assume $h_{i,K}$, and hence $(h_{i,K})_+$, is $L_{i,K}^H(\mathcal{D})$-Lipschitz in (θ_g, θ_i):

$$\big|h_{i,K}(\theta_g, \theta_i) - h_{i,K}(\theta_g', \theta_i')\big| \leq L_{i,K}^H(\mathcal{D}) \left\| \begin{bmatrix} \theta_g - \theta_g' \\ \theta_i - \theta_i' \end{bmatrix} \right\|, \quad \forall (\theta_g, \theta_i), (\theta_g', \theta_i'). \tag{21}$$

Assumption 3.6 (**Personalized posterior concentration**). For each fixed old history, assume $f_{i,K}$ is $\sigma^2_{F,i,K}$-sub-Gaussian under $q_{i,K-1}$, with $\sigma^2_{F,i,K} \le (L^F_{i,K})^2/\tau^F_{i,K}$. For Gaussian Laplace marginals, Assumption 3.4 ensures this condition if

$$\Lambda_{i,K-1} := \left([\Lambda^{-1}_{K-1}]_{ii}\right)^{-1}, \qquad \lambda_{\min}(\Lambda_{i,K-1}) \ge \tau^F_{i,K} > 0.$$

Here $[\Lambda^{-1}_{K-1}]_{ii}$ is the old marginal covariance of θ_i.

Assumption 3.7 (**Decoupled posterior concentration**). For each fixed $\theta_{-i} = \theta^*_{-i}$, let

$$P^{\theta^*_{-i}}_{i,K} := p(\theta_g \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}) p(\theta_i \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}).$$

For almost every $\mathcal{D}$, assume $Q^{\theta^*_{-i}}_{i,K} \ll P^{\theta^*_{-i}}_{i,K}$ and $(h_{i,K})_+$ is $\sigma^2_{H,i,K}(\mathcal{D})$-sub-Gaussian under $P^{\theta^*_{-i}}_{i,K}$, uniformly in θ^*_{-i}, with $\mathbb{E}_{\mathcal{D}}[\sigma^2_{H,i,K}(\mathcal{D})] < \infty$. Appendices A.2 and F.5 give the Gaussian sufficient conditions and reference construction.

Theorem 3.8. *Under Assumptions 3.4 and 3.6, if $q_{i,K} \ll q_{i,K-1}$, the KL bound is*

$$F_{i,K} \le \sqrt{2\sigma^2_{F,i,K} \mathrm{KL}(q_{i,K} \| q_{i,K-1})}. \tag{22}$$

For a fixed old history, averaging over the new all-client task data gives

$$\mathbb{E}_{\mathcal{D}_{1:N,K} | \mathcal{D}_{1:N,1:K-1}}[F_{i,K}] \le \sqrt{2\sigma^2_{F,i,K} \mathbb{E}_{\mathcal{D}_{1:N,K} | \mathcal{D}_{1:N,1:K-1}}[\mathrm{KL}(q_{i,K} \| q_{i,K-1})]}.$$

For exact Bayesian posteriors and their predictive data law, the averaged KL is the information gain at that history. If the same concentration constants hold uniformly over old histories, further averaging yields the population bound

$$\mathbb{E}_{\mathcal{D}_{1:N,1:K}}[F_{i,K}] \le \sqrt{\frac{2(L^F_{i,K})^2}{\tau^F_{i,K}} \mathcal{I}^{\mathrm{intra}}_{\theta_i;[N],K\to<K}}. \tag{23}$$

The KL bounds also hold for Gaussian proxies when both $F_{i,K}$ and the KL use those proxies; their MI interpretation requires Bayesian consistency. Under Assumptions 3.5 and 3.7, with integrable $H_{i,K}, B_{i,K}$ and finite CMI, the population heterogeneity bound is

$$\mathbb{E}_{\mathcal{D}}[H_{i,K}] \le \mathbb{E}_{\mathcal{D}}[B_{i,K}] + \sqrt{2\mathbb{E}_{\mathcal{D}}[\sigma^2_{H,i,K}(\mathcal{D})] \mathcal{I}^{\mathrm{inter}}_{\theta_i \leftrightarrow \theta_g; \mathcal{D}_{1:N,1:K}}}, \tag{24}$$

where the decoupled-posterior baseline is

$$B_{i,K} := \mathbb{E}_{\theta_{-i} | \mathcal{D}_{1:N,1:K}} \left[\mathbb{E}_{P^{\theta_{-i}}_{i,K}} \left[\left(h_{i,K}(\theta_g, \theta_i)\right)_+ \right] \right].$$

The expectations and population CMI refer to the same joint law of task data and parameters.

A KL change-of-measure bound, posterior averaging, and Cauchy–Schwarz over data yield Eqn. (24); the proof also bounds $\mathbb{E}_{\mathcal{D}}[|H_{i,K} - B_{i,K}|]$ by the same square-root term. Thus, population inter-model interference controls the data-averaged deviation from independent pairing of the same conditional marginals.

3.3 Bayesian Personalized Federated Continual Learning (Bayes-pFCL)

Theorem 3.8 motivates directly penalizing the intra- and inter-model interference in local training. Thus, we add intra- and inter-model interference as regularization terms on the vanilla empirical objective in Eqn. 1 to formulate the new empirical loss, *i.e.*,

$$\min_{\{\theta_i\}_{i=1}^N} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \left[\mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}) + \lambda_{\text{intra}} \mathcal{I}^{\mathrm{intra}}_{\theta_i; \mathcal{D}_{1:N,k} \to \mathcal{D}_{1:N,1:k-1}} + \lambda_{\text{inter}} \mathcal{I}^{\mathrm{inter}}_{\theta_i \leftrightarrow \theta_g; \mathcal{D}_{1:N,1:k}} \right], \tag{25}$$

where $\lambda_{\mathrm{intra}}, \lambda_{\mathrm{inter}} \ge 0$ balance the vanilla empirical loss and regularization terms.

The two regularizers target task-induced posterior drift and global–personalized posterior dependence at the population level; Algorithm 1 evaluates their realized Gaussian surrogates. Appendix G.2 discusses the scope of their theoretical guarantees. The dynamic variant of Bayes-pFCL is achieved when using interference defined in Eqns. 13 and 14 in Eqn. (25). When $\lambda_{\mathrm{intra}} = \lambda_{\mathrm{inter}} = 0$, Eqn. (25) reduces to the original local objective.

4 Experiments

Following prior setups [38, 56], we compare Bayes-pFCL with SOTA methods on the following synthetic and real-world datasets. Due to space limitations, we defer the ablation study, sensitivity tests, and additional experimental results to Appendix B.

4.1 Datasets

We evaluate Bayes-pFCL on three image classification datasets: ***EMNIST-Letters*** [8] (26 handwritten letter classes, 28×28 grayscale images), ***CIFAR-100*** [25] (100 fine-grained classes, 32×32 RGB images), and ***TinyImageNet*** [26] (200 classes, 64×64 RGB images). We also use ***TCGA-BRCA*** [49], a breast cancer dataset with 1,095 samples containing over 18,000 gene-expression features, 39 clinical features, and binary subtype labels (Basal-like vs. non-Basal).

4.2 Baselines and Implementation

We compare Bayes-pFCL and its dynamic variant with baselines as follows: FedAvg [37], FedMGP [57], pFedDIL [31] and FedAGC [59], and DCFCL [55]. All SOTA methods are regularization-based pFCL baselines. They use auxiliary signals, such as prompts [57], to construct regularization terms. Additional comparisons with FedAvg + EWC and our reimplementations of pFedDIL [31] and FedAGC [59], together with the evaluation settings and reimplementation details, are provided in Appendix B.3.

All methods are implemented in PyTorch [40] using the same neural network architecture for components with identical roles. Method-specific hyper-parameters are set following the original configurations of the respective baselines.

4.3 Evaluation Metrics

We use *accuracy* and *forgetting* [38] to evaluate classification on *EMNIST-Letters*, *CIFAR-100*, *TinyImageNet*, and *TCGA-BRCA*.

Specifically, let $a_{i,k}$ denote the test accuracy of client i's personalized model θ_i on task k, where θ_i is the parameter after learning all tasks, and let $\bar{a}_{i,k}$ denote the best test accuracy previously achieved on task k before learning the full task sequence. We have

$$\text{Accuracy} := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K} a_{i,k}, \qquad \text{Forgetting} := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K-1}\sum_{k=1}^{K-1}\left(\bar{a}_{i,k} - a_{i,k}\right). \tag{26}$$

4.4 Continual Learning Configuration

Each of the N clients learns K tasks sequentially, with C classes per task. For *EMNIST-Letters*, *CIFAR-100*, and *TinyImageNet*, we follow the standard class-incremental learning (CIL) setting [51], where tasks within each client have disjoint class sets. *TCGA-BRCA* uses four binary classification tasks per client. Appendix H reports the (N, K, C) configuration and training hyperparameters for each dataset.

4.5 Results

EMNIST-Letters. On *EMNIST-Letters*, Bayes-pFCL achieves the strongest classification performance. Bayes-pFCL (Dynamic) obtains the best accuracy of 83.01%, while Bayes-pFCL (Static) achieves the second-best accuracy of 79.34%. Compared with the strongest baseline, FedMGP, Bayes-pFCL (Dynamic) improves accuracy by 9.59%. Meanwhile, its forgetting remains comparable to existing methods, indicating that the proposed interference regularization improves plasticity without substantially sacrificing stability.

CIFAR-100. On *CIFAR-100*, Bayes-pFCL shows a more significant advantage. Bayes-pFCL (Dynamic) achieves the best accuracy of 72.51% and the lowest forgetting of 2.24%, while Bayes-pFCL (Static) obtains the second-best accuracy of 68.92%. Compared with FedMGP, the strongest baseline, Bayes-pFCL (Dynamic) improves accuracy by 26.43% and reduces forgetting by 1.78%. These

Table 1: Efficiency of Bayes-pFCL on training datasets (best result in **bold**, second-best underlined). For classification tasks on *EMNIST-Letters*, *CIFAR-100*, *TinyImageNet*, and *TCGA-BRCA*, we report *Accuracy(ACC)* (%) and *Forgetting* (%). Additional baseline comparisons and reimplementation details are reported in Appendix B.3. "Bayes-pFCL (Static/Dynamic)" refer to Bayes-pFCL with interference-based regularization terms defined under static and dynamic Bayesian frameworks, respectively

Method	**EMNIST-Letters**		**CIFAR-100**		**TinyImageNet**		**TCGA-BRCA**	
	ACC (%)↑	Forgetting (%)↓	ACC (%)↑	Forgetting (%)↓	ACC (%)↑	Forgetting (%)↓	ACC (%)↑	Forgetting (%)↓
FedAvg	65.82 ± 3.47	4.70± 2.57	25.16 ± 2.65	3.76 ± 1.86	34.84 ± 3.71	4.51 ± 3.01	45.37 ± 3.81	3.49 ± 1.48
FedMGP	73.41 ± 3.03	4.88 ± 2.48	31.19 ± 2.81	4.02 ± 1.53	29.80 ± 2.54	5.12 ± 1.82	52.73 ± 1.56	3.71 ± 1.09
DCFCL	68.84 ± 2.14	3.40± 2.90	46.08 ± 2.70	2.73 ± 1.99	41.72 ± 1.79	**3.12** ± 1.98	59.74 ± 1.97	**2.22** ± 1.46
Bayes-pFCL (Static)	79.34 ± 3.17	3.46 ± 1.26	68.92 ± 2.55	2.42 ± 1.56	70.17 ± 3.81	4.20 ± 2.03	73.84 ± 2.60	3.02 ± 1.42
Bayes-pFCL (Dynamic)	**83.01** ± 3.24	**3.34** ± 1.21	**72.51** ± 2.28	**2.24** ± 1.45	**73.09** ± 4.39	5.57 ± 1.88	**75.67** ± 2.48	2.93 ± 1.25

results suggest that Bayes-pFCL can effectively mitigate both data heterogeneity and catastrophic forgetting on more complex classification tasks.

TinyImageNet. On *TinyImageNet*, Bayes-pFCL also clearly outperforms all baselines in accuracy. Bayes-pFCL (Dynamic) achieves the best accuracy of 73.09%, followed by Bayes-pFCL (Static) with 70.17%. Compared with the strongest baseline, FedMGP, Bayes-pFCL (Dynamic) improves accuracy by 31.37%. However, its forgetting is higher than that of several baselines, indicating a stronger stability–plasticity trade-off on this more challenging visual dataset. This suggests that the dynamic Bayesian formulation encourages stronger adaptation to new tasks, but may require further control to reduce forgetting on high-complexity image benchmarks.

TCGA-BRCA. On *TCGA-BRCA*, Bayes-pFCL again achieves the best overall performance. Bayes-pFCL (Dynamic) obtains the highest accuracy of 75.67% and the second-lowest forgetting of 2.93%, while Bayes-pFCL (Static) achieves the second-best accuracy of 73.84%. These results show that Bayes-pFCL is also effective beyond standard image benchmarks, demonstrating its applicability to biomedical classification tasks with heterogeneous clients.

5 Discussion

Bayes-pFCL supports collaborative personalized learning without exchanging raw client data, which may benefit applications with heterogeneous populations and evolving data. However, shared model updates and posterior summaries can still leak information, and heterogeneous data may amplify client-specific biases. High-stakes deployment should therefore be accompanied by privacy protection, fairness assessment, calibration, and domain-specific validation. Appendix G discusses these potential social impacts in more detail.

6 Limitation

The static version of Bayes-pFCL requires task-level Hessian estimation to update precision blocks, incurring $\mathcal{O}(d_i^2)$ memory and $\mathcal{O}(d_i C_{\nabla_i})$ computation per client, where d_i is the parameter dimension and C_{∇_i} is the gradient evaluation cost. Fisher-based, diagonal, block-diagonal, Kronecker-factored, or low-rank approximations [23, 35, 43, 9] reduce overhead at the cost of posterior fidelity. The dynamic version incurs additional overhead by updating precision blocks every communication round rather than after each task.

7 Conclusion

We propose Bayes-pFCL, a Bayesian framework for personalized federated continual learning that represents knowledge as posterior belief. Its intra- and inter-model interference measures quantify task-induced posterior drift and global–personalized coupling through conditional mutual information. Bounds on expected forgetting loss and the expected excess positive personalized–global loss gap over a decoupled-posterior baseline motivate an interference-regularized local objective.

Acknowledgement

We thank the anonymous reviewers and the area chair for their valuable feedback. The work of Q. Yu and H. Wang at Stevens Institute of Technology was supported in part by the United States National Science Foundation (NSF) under grants 2523997, 2315612, and 2332638 and by the AWS Cloud Credit for Research program. The work of H. Wang at the University of Illinois Urbana-Champaign was supported in part by the Amazon Faculty Research Award, Microsoft AI & Society Fellowship, Nokia Faculty Research Award, NSF CAREER Award IIS-2340125, NIH grant R01CA297832, and NSF grant IIS-2127918. The work of Y. Ning was supported in part by NSF under grants 2047843 and 2437621. The work of Q. Zhang was supported in part by the Natural Sciences and Engineering Research Council of Canada ALLRP 588144 - 23. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Wuerkaixi Abudukelimu, Cui Sen, Zhang Jingfeng, Yan Kunda, Han Bo, Niu Gang, Fang Lei, Zhang Changshui, and Sugiyama Masashi. Accurate forgetting for heterogeneous federated continual learning. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[2] John R Anderson. Retrieval of propositional information from long-term memory. *Cognitive psychology*, 6(4):451–474, 1974.

[3] John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

[4] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 111(4):1036, 2004.

[5] Zarka Bashir, Mridula Verma, and C. Krishna Mohan. Hybrid federated continual graph contrastive learning for evolving money laundering threats. *Data Mining and Knowledge Discovery*, 40(2):19, 2026. doi: 10.1007/s10618-026-01192-0. URL `https://doi.org/10.1007/s10618-026-01192-0`.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] Fernando E Casado, Dylan Lema, Roberto Iglesias, Carlos V Regueiro, and Senén Barro. Federated and continual learning for classification tasks in a society of devices. *arXiv preprint arXiv:2006.07129*, 2020. URL `https://arxiv.org/abs/2006.07129`.

[8] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

[9] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux: Effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, 2021.

[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3366–3385, 2021. doi: 10.1109/TPAMI.2021.3057446.

[11] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10154–10163. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00992.

[12] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. pfl/personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568, 2020.

[13] Qi Fan, Zhaoxia Li, Haifeng Chen, Jinchun Liu, Yang Xin, Bin Jia, and Jeng-Shyang Pan. High-performance personalized federated continual learning framework for dynamic data sharing in iov. *IEEE Internet of Things Journal*, 13(8):15265–15279, 2026.

[14] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, FL, 3 edition, 2013.

[15] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, and Huiping Zhuang. AFL: A single-round analytic approach for federated learing with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[18] Sean M. Hendryx, Dharma Raj KC, Bradley Walls, and Clayton T. Morrison. Federated reconnaissance: Efficient, distributed, class-incremental learning. *arXiv preprint arXiv:2109.00150*, 2021. URL `https://arxiv.org/abs/2109.00150`.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[20] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10143–10153, 2022. doi: 10.1109/CVPR52688.2022.00990.

[21] K. Jagdeesh, N. Kanimozhi, Tanvir H. Sardar, N. Naveenkumar, B. Mahalakshmi, A. Chandrasekar, M. Karpagam, and Sk Mahmudul Hasan. Federated learning with continual update for privacy-preserving clinical event prediction across distributed hospitals using MCN-GNN. *Scientific Reports*, 16(1):12608, 2026. doi: 10.1038/s41598-026-40964-y.

[22] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith 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. doi: 10.1561/2200000083.

[23] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *arXiv*, 2016. doi: 10.1073/pnas.1611835114.

[24] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

[25] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`. Technical Report.

[26] Ya Le. Tiny imagenet visual recognition challenge. `https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf`, 2015.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Baoxue Li, Pengyu Song, Chunhui Zhao, and Min Xie. Facing spatiotemporal heterogeneity: A unified federated continual learning framework with self-challenge rehearsal for industrial monitoring tasks. *Knowledge-Based Systems*, 289:111491, 2024. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2024.111491. URL `https://www.sciencedirect.com/science/article/pii/S0950705124001266`.

[29] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. doi: 10.1109/MSP.2020.2975749.

[30] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[31] Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part VIII*, pages 127–144. Springer, 2024. doi: 10.1007/978-3-031-72952-2_8.

[32] Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2182–2188. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/303. URL `https://doi.org/10.24963/ijcai.2022/303`.

[33] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

[34] Disha Makhija, Joydeep Ghosh, and Nhat Ho. A bayesian approach for personalized federated learning in heterogeneous settings. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102428–102455. Curran Associates, Inc., 2024. doi: 10.52202/079017-3252. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/b9a159b81a0535e3944ffdbd07ae3ba1-Paper-Conference.pdf`.

[35] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2408–2417, 2015.

[36] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. doi: 10.1016/S0079-7421(08)60536-8.

[37] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017.

[38] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, May 2021. URL `https://openreview.net/forum?id=qxEdMZVUqf0`. OpenReview.net.

[39] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[41] Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=cRxYWKiTan`.

[42] Jeroen GW Raaijmakers and Richard M Shiffrin. Search of associative memory. *Psychological review*, 88(2):93, 1981.

[43] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

[44] Babakniya Sara, Fabian Zalan, Chaoyang He, Soltanolkotabi Mahdi, and Avestimehr Salman. A data-free approach to mitigate catastrophic forgetting in federated learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[45] Marco Schreyer, Hamed Hemati, Damian Borth, and Miklos A. Vasarhelyi. Federated continual learning to detect accounting anomalies in financial auditing, 2022. URL https://arxiv.org/abs/2210.15051.

[46] Jonathan A. Scott, Hossein Zakerinia, and Christoph H. Lampert. PeFLL: Personalized federated learning by learning to learn. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MrYiwlDRQO.

[47] Richard M Shiffrin and Mark Steyvers. A model for recognition memory: Rem—retrieving effectively from memory. *Psychonomic bulletin & review*, 4(2):145–166, 1997.

[48] Anay Sinhal, Arpana Sinhal, and Amit Sinhal. Federated continual learning for privacy-preserving hospital imaging classification, 2026.

[49] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

[50] Anastasiia Usmanova, François Portet, Philippe Lalanda, and Germán Vega. A distillation-based approach integrating continual learning and federated learning for pervasive services. *CoRR*, abs/2109.04197, 2021. URL https://arxiv.org/abs/2109.04197.

[51] Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. URL https://arxiv.org/abs/1904.07734.

[52] Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A framework and some existing methods. *TDKE*, 28(12):3395–3408, 2016.

[53] Hao Wang and Dit-Yan Yeung. A survey on bayesian deep learning. *CSUR*, 53(5):1–37, 2020.

[54] Zihao Xu, Guangyuan Hao, Hao He, and Hao Wang. Domain indexing variational bayes: Interpretable domain index for domain adaptation. In *ICLR*, 2023.

[55] Danni Yang, Zhikang Chen, Sen Cui, Mengyue Yang, Ding Li, Abudukelimu Wuerkaixi, Haoxuan Li, Jinke Ren, and Mingming Gong. Decentralized dynamic cooperation of personalized models for federated continual learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.23683.

[56] Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 2021.

[57] Hao Yu, Xin Yang, Xin Gao, Yan Kang, Hao Wang, Junbo Zhang, and Tianrui Li. Personalized federated continual learning via multi-granularity prompt. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3637528.3671948. URL https://doi.org/10.1145/3637528.3671948.

[58] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Trong Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7252–7261. PMLR, 2019.

[59] Chengchao Zhang, Fanhua Shang, Hongying Liu, Liang Wan, and Wei Feng. FedAGC: Federated continual learning with asymmetric gradient correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[60] Jianqing Zhang, Yang Hua, Haibing Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–11244, 2023.

[61] Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26293–26310. PMLR, 2022.

[62] Siwei Zheng, Jia Hu, Geyong Min, and Ke Li. Meta representation-based personalized federated continual learning in edge computing systems. *IEEE Transactions on Computational Social Systems*, 2025. URL https://ieeexplore.ieee.org/document/11081891/.

[63] Tianshuo Zhou and Boyuan Wang. Cross paradigm fusion of federated and continual learning on multilayer perceptron mixer architecture for incremental thoracic infection diagnosis. *Scientific Reports*, 15(1):24449, 2025. doi: 10.1038/s41598-025-06077-8. URL https://doi.org/10.1038/s41598-025-06077-8.

A Theoretical Analysis

A.1 Theorem 3.3

Proof. By the Laplace approximation, the precision matrix is the Hessian of the negative log-posterior at the MAP point:

$$\Lambda_K = \nabla_z^2 \left[-\log p(z \mid \mathcal{D}_{1:N,1:K}) \right]_{z=\mu_K}.$$

From Eqn. (6),

$$-\log p(z \mid \mathcal{D}_{1:N,1:K}) = -\log p(\theta_g) - \sum_{j=1}^{N} \log p(\theta_j \mid \theta_g) - \sum_{j=1}^{N} \sum_{k=1}^{K} \log p(\mathcal{D}_{j,k} \mid \theta_g, \theta_j) + C.$$

Taking the mixed derivative with respect to θ_g and θ_i, the global prior $p(\theta_g)$ contributes nothing. The terms $p(\theta_j \mid \theta_g)$ and $p(\mathcal{D}_{j,k} \mid \theta_g, \theta_j)$ for $j \neq i$ do not depend on θ_i, and therefore also contribute nothing. Hence,

$$\Lambda_{gi,K} = \nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(\theta_i \mid \theta_g) \right]_{z=\mu_K} + \sum_{k=1}^{K} \nabla_{\theta_g} \nabla_{\theta_i}^\top \left[-\log p(\mathcal{D}_{i,k} \mid \theta_g, \theta_i) \right]_{z=\mu_K}.$$

This proves the claim. □

Interpretation of the online precision recursion. The decomposition above evaluates all prior and likelihood Hessians at the same current MAP. The precision stored at the preceding task is generally evaluated at a different MAP, so the decomposition does not by itself imply an additive update with that stored precision. The recursion following Theorem 3.3 uses the fixed old-Gaussian interpretation derived in Appendix F.1.

A.2 Bounding Forgetting and Heterogeneity via Sub-Gaussian Concentration

In this subsection, we bound forgetting-induced loss and the excess positive personalized–global loss gap over a decoupled-posterior baseline. The argument starts from a KL change-of-measure bound on differences of expectations. Gaussian concentration supplies the required sub-Gaussian conditions for the Laplace proxies. For forgetting, we then distinguish a fixed-history information-gain bound from its population mutual-information corollary; these identities require coherent Bayesian updates. For heterogeneity, we first obtain a fixed-data KL bound and then average over data, allowing a data-dependent concentration constant with finite mean.

Fix a task $K \geq 2$, write $(a)_+ := \max\{a, 0\}$, and let

$$\mathcal{D}_{<K} := \mathcal{D}_{1:N,1:K-1}, \qquad \mathcal{D}_K := \mathcal{D}_{1:N,K}, \qquad \mathcal{D}_{\leq K} := \mathcal{D}_{1:N,1:K}.$$

For client i, the population intra-model interference is

$$\mathcal{I}^{\text{intra}}_{i,K\to<K} := I(\theta_i; \mathcal{D}_K \mid \mathcal{D}_{<K}),$$

which abbreviates the all-client-data intra-model term used in Eqn. (25). Its fixed-history counterpart is defined below by adding the observed history as an argument.

Let $q_{i,K-1}$ and $q_{i,K}$ denote the marginal posteriors of θ_i before and after learning task K, respectively:

$$q_{i,K-1}(\theta_i) := p(\theta_i \mid \mathcal{D}_{<K}), \qquad q_{i,K}(\theta_i) := p(\theta_i \mid \mathcal{D}_{\leq K}).$$

Under the Laplace approximation, we write

$$q_{i,K-1} \approx \mathcal{N}\left(\mu_{i,K-1}, \Lambda_{i,K-1}^{-1}\right), \qquad q_{i,K} \approx \mathcal{N}\left(\mu_{i,K}, \Lambda_{i,K}^{-1}\right),$$

Here the marginal precision is

$$\Lambda_{i,k} := \left([\Lambda_k^{-1}]_{ii}\right)^{-1}, \qquad k \in \{K-1, K\},$$

where Λ_k is the joint precision and $d_i = \dim(\theta_i)$. In general, it differs from the joint diagonal block $\Lambda_{ii,k}$, whose inverse is a conditional covariance. The KL bound below also applies to arbitrary candidate pairs on the same parameter space. When the pair consists of Gaussian proxies, all occurrences of $F_{i,K}$ and its KL divergence in that bound are evaluated under those proxies.

A.2.1 Preliminaries: KL Variational Bound

Lemma A.1 (KL variational bound for sub-Gaussian functions)**.** *Let P and Q be probability measures on a measurable space Θ with $Q \ll P$. Let $f : \Theta \to \mathbb{R}$ be σ^2-sub-Gaussian under P, i.e.,*

$$\mathbb{E}_P[\exp(\lambda(f - \mathbb{E}_P[f]))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \qquad \forall \lambda \in \mathbb{R}.$$

Then

$$|\mathbb{E}_Q[f] - \mathbb{E}_P[f]| \leq \sqrt{2\sigma^2 \mathrm{KL}(Q\|P)}.$$

Proof. By the Donsker–Varadhan variational representation of KL divergence, for any $\lambda \in \mathbb{R}$,

$$\lambda(\mathbb{E}_Q[f] - \mathbb{E}_P[f]) - \log \mathbb{E}_P[\exp(\lambda(f - \mathbb{E}_P[f]))] \leq \mathrm{KL}(Q\|P).$$

Using the sub-Gaussian MGF bound gives

$$\lambda(\mathbb{E}_Q[f] - \mathbb{E}_P[f]) \leq \mathrm{KL}(Q\|P) + \frac{\lambda^2\sigma^2}{2}.$$

Optimizing over $\lambda > 0$ yields

$$\mathbb{E}_Q[f] - \mathbb{E}_P[f] \leq \sqrt{2\sigma^2 \mathrm{KL}(Q\|P)}.$$

Applying the same argument to $-f$ gives the two-sided bound. □

The direction of the KL divergence is important: f must be sub-Gaussian under the reference distribution P, which appears in the second argument of $\mathrm{KL}(Q\|P)$.

A.2.2 Sub-Gaussianity via the Laplace Approximation

Assumption A.2 (Lipschitz losses)**.** For client i, define the averaged old-task personalized loss

$$f_{i,K}(\theta_i) := \frac{1}{K-1} \sum_{k=1}^{K-1} \mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}).$$

Assume that $f_{i,K}$ is $L^F_{i,K}$-Lipschitz in θ_i.

Define the old-task personalized–global loss gap

$$h_{i,K}(\theta_g, \theta_i) := \frac{1}{K-1} \sum_{k=1}^{K-1} \left[\mathcal{L}_i(\theta_i; \mathcal{D}_{i,k}) - \mathcal{L}_i(\theta_g; \mathcal{D}_{i,k})\right].$$

For each observed history $\mathcal{D}_{\leq K}$, assume that $h_{i,K}$ is $L^H_{i,K}(\mathcal{D}_{\leq K})$-Lipschitz in (θ_g, θ_i). This constant may depend on the data.

For an untruncated Gaussian Laplace posterior, these Lipschitz conditions are required globally on its support. Their applicability to practical training is discussed in Appendix G.2.

Assumption A.3 (Posterior precision lower bounds)**.** For the Gaussian proxy analysis, assume that the old marginal approximation is

$$q_{i,K-1} = \mathcal{N}\left(\mu_{i,K-1}, \Lambda^{-1}_{i,K-1}\right),$$

where

$$\lambda_{\min}(\Lambda_{i,K-1}) \geq \tau^F_{i,K} > 0.$$

This precision condition guarantees concentration for the Gaussian proxy; it does not by itself guarantee concentration for the exact old posterior.

Moreover, for each fixed value θ^*_{-i} of $\theta_{-i} := \{\theta_j\}_{j \neq i}$, define

$$Q^{\theta^*_{-i}}_{i,K} := p(\theta_g, \theta_i \mid \theta_{-i} = \theta^*_{-i}, \mathcal{D}_{\leq K}),$$

and

$$P^{\theta^*_{-i}}_{i,K} := p(\theta_g \mid \theta_{-i} = \theta^*_{-i}, \mathcal{D}_{\leq K}) p(\theta_i \mid \theta_{-i} = \theta^*_{-i}, \mathcal{D}_{\leq K}).$$

For the Gaussian sufficient conditions for heterogeneity, all posterior expectations and conditionals at each observed history are evaluated under the same joint Gaussian approximation. In particular, $P_{i,K}^{\theta^*_{-i}}$ is the product of the two conditional marginals of $Q_{i,K}^{\theta^*_{-i}}$, so that

$$P_{i,K}^{\theta^*_{-i}} = \mathcal{N}\Big(\mu_{P_{i,K}}^{\theta^*_{-i}}, (\Lambda_{P_{i,K}}^{\theta^*_{-i}})^{-1}\Big),$$

with the precision lower bound

$$\lambda_{\min}(\Lambda_{P_{i,K}}^{\theta^*_{-i}}) \geq \tau_{i,K}^H(\mathcal{D}_{\leq K}) > 0, \qquad \forall \theta^*_{-i}.$$

The lower bound is uniform over θ^*_{-i} at each fixed history, but it may vary with $\mathcal{D}_{\leq K}$.

Lemma A.4 (Gaussian concentration of Lipschitz functions)**.** *Let $X \sim \mathcal{N}(\mu, \Sigma)$, and let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then $f(X)$ is $L^2\lambda_{\max}(\Sigma)$-sub-Gaussian:*

$$\mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(X)]))] \leq \exp\left(\frac{\lambda^2 L^2 \lambda_{\max}(\Sigma)}{2}\right), \qquad \forall \lambda \in \mathbb{R}.$$

Proposition A.5 (Sub-Gaussianity of old-task loss and loss gap)**.** *For the Gaussian approximations in Assumptions A.2 and A.3, the following hold:*

1. *The averaged old-task loss $f_{i,K}$ is $\sigma_{F,i,K}^2$-sub-Gaussian under $q_{i,K-1}$, with*
$$\sigma_{F,i,K}^2 \leq \frac{(L_{i,K}^F)^2}{\tau_{i,K}^F}.$$
2. *For each observed history, the positive loss gap $(h_{i,K})_+$ is $\sigma_{H,i,K}^2(\mathcal{D}_{\leq K})$-sub-Gaussian under $P_{i,K}^{\theta^*_{-i}}$, uniformly in θ^*_{-i}, with*
$$\sigma_{H,i,K}^2(\mathcal{D}_{\leq K}) \leq \frac{\big(L_{i,K}^H(\mathcal{D}_{\leq K})\big)^2}{\tau_{i,K}^H(\mathcal{D}_{\leq K})}.$$

Proof. For the first claim, the Gaussian old marginal gives

$$\theta_i \sim \mathcal{N}\Big(\mu_{i,K-1}, \Lambda_{i,K-1}^{-1}\Big) \quad \text{under } q_{i,K-1}.$$

Since $f_{i,K}$ is $L_{i,K}^F$-Lipschitz, Lemma A.4 implies that $f_{i,K}$ is

$$(L_{i,K}^F)^2 \lambda_{\max}(\Lambda_{i,K-1}^{-1})$$

sub-Gaussian. By the precision lower bound,

$$\lambda_{\max}(\Lambda_{i,K-1}^{-1}) = \frac{1}{\lambda_{\min}(\Lambda_{i,K-1})} \leq \frac{1}{\tau_{i,K}^F}.$$

Thus,

$$\sigma_{F,i,K}^2 \leq \frac{(L_{i,K}^F)^2}{\tau_{i,K}^F}.$$

The second claim follows similarly. Under the same Gaussian approximation, for each fixed θ^*_{-i},

$$P_{i,K}^{\theta^*_{-i}} = \mathcal{N}\Big(\mu_{P_{i,K}}^{\theta^*_{-i}}, (\Lambda_{P_{i,K}}^{\theta^*_{-i}})^{-1}\Big).$$

Since the positive-part map is 1-Lipschitz, $(h_{i,K})_+$ is also $L_{i,K}^H(\mathcal{D}_{\leq K})$-Lipschitz. Lemma A.4 therefore implies that $(h_{i,K})_+$ is

$$\big(L_{i,K}^H(\mathcal{D}_{\leq K})\big)^2 \lambda_{\max}\Big((\Lambda_{P_{i,K}}^{\theta^*_{-i}})^{-1}\Big)$$

sub-Gaussian under $P_{i,K}^{\theta^*_{-i}}$. The uniform lower bound $\lambda_{\min}(\Lambda_{P_{i,K}}^{\theta^*_{-i}}) \geq \tau_{i,K}^H(\mathcal{D}_{\leq K})$ gives

$$\lambda_{\max}\Big((\Lambda_{P_{i,K}}^{\theta^*_{-i}})^{-1}\Big) \leq \frac{1}{\tau_{i,K}^H(\mathcal{D}_{\leq K})}.$$

Therefore,

$$\sigma_{H,i,K}^2(\mathcal{D}_{\leq K}) \leq \frac{\big(L_{i,K}^H(\mathcal{D}_{\leq K})\big)^2}{\tau_{i,K}^H(\mathcal{D}_{\leq K})}.$$

□

A.2.3 Bound on Loss Induced by Catastrophic Forgetting

Although the posterior incorporates all clients' data, $f_{i,K}$ still evaluates only client i's previous tasks. Define its positive posterior expected loss increase by

$$F_{i,K} := \left(\mathbb{E}_{q_{i,K}}[f_{i,K}(\theta_i)] - \mathbb{E}_{q_{i,K-1}}[f_{i,K}(\theta_i)]\right)_+ .$$

Theorem A.6 (KL bound on forgetting)**.** *Fix the observed all-client history $\mathcal{D}_{<K} = \mathcal{D}^*_{<K}$. Below, conditioning on $\mathcal{D}^*_{<K}$ abbreviates conditioning on this fixed value of $\mathcal{D}_{<K}$. Let $q_{i,K-1}$ be the reference distribution and $q_{i,K}$ any candidate updated distribution on the same parameter space. Assume $q_{i,K} \ll q_{i,K-1}$, finite KL divergence, and that $f_{i,K}$ is $\sigma^2_{F,i,K}$-sub-Gaussian under $q_{i,K-1}$, with*

$$\sigma^2_{F,i,K} \le \frac{(L^F_{i,K})^2}{\tau^F_{i,K}}.$$

Then

$$F_{i,K} \le \sqrt{2\sigma^2_{F,i,K}\mathrm{KL}(q_{i,K}\|q_{i,K-1})}. \tag{27}$$

If the same reference and sub-Gaussian constant apply to the candidates indexed by new data $\mathcal{D}_K$, with finite expected KL, then under any specified conditional law of those data,

$$\mathbb{E}_{\mathcal{D}_K|\mathcal{D}^*_{<K}}[F_{i,K}] \le \sqrt{2\sigma^2_{F,i,K}\mathbb{E}_{\mathcal{D}_K|\mathcal{D}^*_{<K}}[\mathrm{KL}(q_{i,K}\|q_{i,K-1})]}. \tag{28}$$

The constants may depend on the fixed history, but not on the new data.

Proof. The positive part is bounded by the absolute difference of expectations. Apply Lemma A.1 with $P = q_{i,K-1}$, $Q = q_{i,K}$, and $f = f_{i,K}$, using the assumed sub-Gaussianity under the old reference distribution. This proves Eqn. (27). Taking the conditional expectation over the new data and applying Jensen's inequality proves Eqn. (28). □

The reference concentration assumption concerns the distributions actually used in this theorem. Proposition A.5 supplies it for the Gaussian proxies. Applying the theorem to exact posteriors instead requires the same sub-Gaussian condition to hold for the exact old posterior.

Corollary A.7 (Information gain for a fixed history)**.** *Under Theorem A.6, suppose additionally that $q_{i,K-1}$ and the family $q_{i,K}$ are exact marginal posteriors from the same Bayesian model, and average new data under its posterior-predictive law $p(\mathcal{D}_K \mid \mathcal{D}^*_{<K})$. Define the history-conditioned intra-model information gain by*

$$\begin{aligned}\mathcal{I}^{\mathrm{intra}}_{i,K\to<K}(\mathcal{D}^*_{<K}) &:= \mathbb{E}_{\mathcal{D}_K|\mathcal{D}^*_{<K}}[\mathrm{KL}(q_{i,K}\|q_{i,K-1})] \\ &= I(\theta_i; \mathcal{D}_K \mid \mathcal{D}_{<K} = \mathcal{D}^*_{<K}).\end{aligned} \tag{29}$$

Then

$$\mathbb{E}_{\mathcal{D}_K|\mathcal{D}^*_{<K}}[F_{i,K}] \le \sqrt{\frac{2(L^F_{i,K})^2}{\tau^F_{i,K}}\mathcal{I}^{\mathrm{intra}}_{i,K\to<K}(\mathcal{D}^*_{<K})}. \tag{30}$$

Proof. For the coherent Bayesian model, the integrand in the expected KL is the posterior expectation of $\log[p(\theta_i \mid \mathcal{D}^*_{<K}, \mathcal{D}_K)/p(\theta_i \mid \mathcal{D}^*_{<K})]$. Posterior-predictive averaging gives the conditional joint-vs-product definition of mutual information in Eqn. (29). Substituting this identity into Eqn. (28) and using the assumed variance proxy proves the result. □

Corollary A.8 (Population mutual-information bound)**.** *Suppose Corollary A.7 applies for almost every old history under the same Bayesian model, and the variance proxy $(L^F_{i,K})^2/\tau^F_{i,K}$ is an upper bound independent of that history. Then*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{<K}}\left[\mathcal{I}^{\mathrm{intra}}_{i,K\to<K}(\mathcal{D}_{<K})\right] &= \mathcal{I}^{\mathrm{intra}}_{i,K\to<K} \\ &= I(\theta_i; \mathcal{D}_K \mid \mathcal{D}_{<K}),\end{aligned} \tag{31}$$

and

$$\mathbb{E}_{\mathcal{D}_{\le K}}[F_{i,K}] \le \sqrt{\frac{2(L^F_{i,K})^2}{\tau^F_{i,K}}\mathcal{I}^{\mathrm{intra}}_{i,K\to<K}}. \tag{32}$$

All data expectations use the joint law of that Bayesian model.

Proof. Conditional mutual information averages the fixed-history information gain, giving Eqn. (31). Average Eqn. (30) over the old history and apply Jensen's inequality with the uniform variance proxy. □

Scope of the Gaussian proxy. Appendix F.2 gives the construction and closed form of the realized Gaussian KL penalty. Under the Gaussian reference conditions, Theorem A.6 bounds the loss increase computed under these proxies. Separately fitting old and new Laplace approximations does not ensure that their family is a coherent Bayesian update. Their expected KL is therefore a computable information-gain surrogate, not automatically the exact population or fixed-history mutual information. The MI corollaries also apply to a coherent approximate Bayesian model only when its prior, predictive law, posterior family, and old-reference concentration condition are mutually consistent.

A.2.4 Bound on the Positive Personalized–Global Loss Gap

Fix a joint law of $(\theta_g, \theta_{1:N}, \mathcal{D}_{\le K})$. At each observed history $\mathcal{D}_{\le K}$, define the expected positive old-task loss gap of client i after task K as

$$H_{i,K} := \mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}}\left[\mathbb{E}_{Q_{i,K}^{\theta_{-i}}}\left[\left(h_{i,K}(\theta_g,\theta_i)\right)_+\right]\right].$$

The positive part is evaluated before posterior averaging, so positive loss gaps are not canceled by negative ones. The dependence of $H_{i,K}$, $h_{i,K}$, and the conditional distributions on the observed data is suppressed in the notation. These quantities vary when the data history varies. For fixed data and θ^*_{-i}, the reference $P_{i,K}^{\theta^*_{-i}}$ is the product of the two conditional marginals of the same $Q_{i,K}^{\theta^*_{-i}}$. The population inter-model interference in Eqn. (10) satisfies

$$\begin{aligned}
&\mathcal{I}^{\mathrm{inter}}_{\theta_i\leftrightarrow\theta_g;\mathcal{D}_{1:N,1:K}} \\
&\quad= \mathbb{E}_{\mathcal{D}_{\le K}}\mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}}\left[\mathrm{KL}\left(Q_{i,K}^{\theta_{-i}}\,\middle\|\,P_{i,K}^{\theta_{-i}}\right)\right] \\
&\quad= I(\theta_g;\theta_i \mid \theta_{-i},\mathcal{D}_{\le K}).
\end{aligned}$$

Both expectations use the chosen joint law. In particular, the conditional mean over θ_{-i} at one observed history is not the population CMI. The corresponding fixed-data baseline is

$$B_{i,K} := \mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}}\left[\mathbb{E}_{P_{i,K}^{\theta_{-i}}}\left[\left(h_{i,K}(\theta_g,\theta_i)\right)_+\right]\right],$$

with its data dependence likewise suppressed.

Theorem A.9 (Population inter-model interference bounds deviation from the decoupled baseline)**.** *For almost every observed history, assume that $Q_{i,K}^{\theta^*_{-i}} \ll P_{i,K}^{\theta^*_{-i}}$ and that $(h_{i,K})_+$ is $\sigma^2_{H,i,K}(\mathcal{D}_{\le K})$-sub-Gaussian under $P_{i,K}^{\theta^*_{-i}}$, uniformly over the conditional values θ^*_{-i}. Assume that $H_{i,K}$ and $B_{i,K}$ are integrable under the data law and that*

$$\mathbb{E}_{\mathcal{D}_{\le K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\right] < \infty, \qquad \mathcal{I}^{\mathrm{inter}}_{\theta_i\leftrightarrow\theta_g;\mathcal{D}_{1:N,1:K}} < \infty.$$

Then the fixed-data KL bound is

$$|H_{i,K} - B_{i,K}| \le \sqrt{2\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}}\left[\mathrm{KL}\left(Q_{i,K}^{\theta_{-i}}\,\middle\|\,P_{i,K}^{\theta_{-i}}\right)\right]}. \tag{33}$$

Averaging over data gives

$$\mathbb{E}_{\mathcal{D}_{\le K}}[|H_{i,K} - B_{i,K}|] \le \sqrt{2\mathbb{E}_{\mathcal{D}_{\le K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\right]\mathcal{I}^{\mathrm{inter}}_{\theta_i\leftrightarrow\theta_g;\mathcal{D}_{1:N,1:K}}}. \tag{34}$$

Consequently,

$$\mathbb{E}_{\mathcal{D}_{\le K}}[H_{i,K}] \le \mathbb{E}_{\mathcal{D}_{\le K}}[B_{i,K}] + \sqrt{2\mathbb{E}_{\mathcal{D}_{\le K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\right]\mathcal{I}^{\mathrm{inter}}_{\theta_i\leftrightarrow\theta_g;\mathcal{D}_{1:N,1:K}}}.$$

Proof. Fix the observed history and $\theta_{-i} = \theta^*_{-i}$. Apply Lemma A.1 with $P = P_{i,K}^{\theta^*_{-i}}$, $Q = Q_{i,K}^{\theta^*_{-i}}$, and $f = (h_{i,K})_+$. The assumed reference concentration gives

$$\left|\mathbb{E}_{Q_{i,K}^{\theta^*_{-i}}}[(h_{i,K})_+] - \mathbb{E}_{P_{i,K}^{\theta^*_{-i}}}[(h_{i,K})_+]\right| \leq \sqrt{2\sigma^2_{H,i,K}(\mathcal{D}_{\leq K})\mathrm{KL}\Big(Q_{i,K}^{\theta^*_{-i}} \,\Big\|\, P_{i,K}^{\theta^*_{-i}}\Big)}.$$

The triangle inequality and Jensen's inequality for the conditional average over $\theta_{-i} \mid \mathcal{D}_{\leq K}$ prove Eqn. (33). The concentration constant is uniform over θ_{-i} at that history, but may depend on the history. Averaging the fixed-data inequality and applying Cauchy–Schwarz yields

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\leq K}}[|H_{i,K} - B_{i,K}|] \\
&\quad\leq \sqrt{2}\,\mathbb{E}_{\mathcal{D}_{\leq K}}\left[\sqrt{\sigma^2_{H,i,K}(\mathcal{D}_{\leq K})}\sqrt{\mathbb{E}_{\theta_{-i}|\mathcal{D}_{\leq K}}\left[\mathrm{KL}\Big(Q_{i,K}^{\theta_{-i}} \,\Big\|\, P_{i,K}^{\theta_{-i}}\Big)\right]}\right] \\
&\quad\leq \sqrt{2\mathbb{E}_{\mathcal{D}_{\leq K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\leq K})\right]\mathbb{E}_{\mathcal{D}_{\leq K}}\mathbb{E}_{\theta_{-i}|\mathcal{D}_{\leq K}}\left[\mathrm{KL}\Big(Q_{i,K}^{\theta_{-i}} \,\Big\|\, P_{i,K}^{\theta_{-i}}\Big)\right]}.
\end{aligned}$$

The conditional joint-vs-product identity, including the outer data expectation, identifies the final expected KL with $I(\theta_g; \theta_i \mid \theta_{-i}, \mathcal{D}_{\leq K})$. This proves Eqn. (34); no independence between the data-dependent variance proxy and the conditional KL is needed. The final bound follows from $H_{i,K} \leq B_{i,K} + |H_{i,K} - B_{i,K}|$ and data averaging. □

For the Gaussian sufficient conditions in Proposition A.5, one may take

$$\sigma^2_{H,i,K}(\mathcal{D}_{\leq K}) \leq \frac{\big(L^H_{i,K}(\mathcal{D}_{\leq K})\big)^2}{\tau^H_{i,K}(\mathcal{D}_{\leq K})}.$$

Finite expectation of the ratio on the right is sufficient for the required finite mean variance proxy. A deterministic bound valid for every data history is not required. For exact posteriors, the sub-Gaussian assumption must hold under their actual decoupled reference distributions. If Gaussian proxies are used instead, the data law and the chosen conditional joint Gaussian family together define the joint law for $H_{i,K}$, $B_{i,K}$, and the population MI in this theorem. All three quantities must use this same law; the proxy MI is not thereby identified with the MI of the original exact posterior model. Appendix G.2 discusses the scope of these conditional guarantees in practice.

The forgetting result first bounds the increase in client i's old-task loss by its posterior KL drift, conditioned on the all-client history. Coherent Bayesian updates identify the predictive expectation of this KL with fixed-history information gain. A uniform reference concentration condition then gives the population bound in terms of $I(\theta_i; \mathcal{D}_K \mid \mathcal{D}_{<K})$.

Second, the population inter-model interference

$$I(\theta_g; \theta_i \mid \theta_{-i}, \mathcal{D}_{\leq K})$$

bounds the data-averaged deviation $\mathbb{E}_{\mathcal{D}_{\leq K}}[|H_{i,K} - B_{i,K}|]$, with a coefficient determined by the mean reference variance proxy. It also controls the expected positive excess $\mathbb{E}_{\mathcal{D}_{\leq K}}[(H_{i,K} - B_{i,K})_+]$. The bound on $\mathbb{E}_{\mathcal{D}_{\leq K}}[H_{i,K}]$ retains $\mathbb{E}_{\mathcal{D}_{\leq K}}[B_{i,K}]$; inter-model interference alone does not control changes in the conditional marginals or their baseline. The population MI cannot replace the conditional expected KL in the fixed-data bound without also averaging the left-hand side over data.

The Gaussian reference conditions justify concentration for the proxy bounds. For intra-model interference, an expected Gaussian KL remains a surrogate unless the Bayesian-coherence conditions above also hold. For inter-model interference, the expected joint-vs-product KL is the population CMI under the specified joint law; the same data averaging is required when this law is constructed from Gaussian proxies.

For the exact fixed-history bound and any $\lambda_{\mathrm{intra}} > 0$, Young's inequality also gives

$$\mathbb{E}_{\mathcal{D}_K|\mathcal{D}^*_{<K}}[F_{i,K}] \leq \lambda_{\mathrm{intra}}\mathcal{I}^{\mathrm{intra}}_{i,K\to<K}(\mathcal{D}^*_{<K}) + \frac{\sigma^2_{F,i,K}}{2\lambda_{\mathrm{intra}}}.$$

Table 2: Ablation study of Bayes-pFCL on *EMNIST-Letters*, *CIFAR-100*, *TinyImageNet*, and *TCGA-BRCA*. We report *Accuracy* (%) and *Forgetting* (%), where higher accuracy and lower forgetting are better. "*w/o* $\mathcal{I}^{\text{intra}}$" and "*w/o* $\mathcal{I}^{\text{inter}}$" denote removing the intra-model and inter-model interference regularizers, respectively.

Method	**EMNIST-Letters**		**CIFAR-100**		**TinyImageNet**		**TCGA-BRCA**	
	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓
Bayes-pFCL (Static)	79.34 ± 3.17	3.46 ± 1.26	68.92 ± 2.55	2.42 ± 1.56	70.17 ± 3.81	4.20 ± 2.03	73.84 ± 2.60	3.02 ± 1.42
w/o $\mathcal{I}^{\text{intra}}$	72.81 ± 2.95	4.07 ± 1.59	61.33 ± 2.13	3.30 ± 1.01	57.38 ± 3.63	2.84 ± 0.75	64.80 ± 2.11	3.21 ± 1.28
w/o $\mathcal{I}^{\text{inter}}$	69.44 ± 2.58	3.11 ± 2.61	58.91 ± 2.39	3.16 ± 2.84	50.49 ± 1.28	3.57 ± 1.20	61.09 ± 1.43	3.88 ± 1.52

Table 3: Ablation of personalization and aggregation on *CIFAR-100*. We report *Accuracy* (%) and *Forgetting* (%), where higher accuracy and lower forgetting are better.

Method	**CIFAR-100**	
	Accuracy (%)↑	Forgetting (%)↓
Bayes-pFCL	68.92 ± 2.55	2.42 ± 1.56
w/o personalization	25.16 ± 2.65	3.76 ± 1.86
w/o aggregation	47.44 ± 2.11	1.89 ± 0.74

The analogous Gaussian-proxy statement uses the expected KL in Eqn. (28) instead of exact information gain. With a uniformly controlled reference variance proxy, this relates the existing intra-model penalty to the forgetting bound.

To relate the population heterogeneity bound to the existing inter-model regularizer, for any $\lambda_{\text{inter}} > 0$, Young's inequality gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\le K}}[H_{i,K}] &\le \mathbb{E}_{\mathcal{D}_{\le K}}[B_{i,K}] + \sqrt{2\mathbb{E}_{\mathcal{D}_{\le K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\right]\mathcal{I}^{\text{inter}}_{\theta_i \leftrightarrow \theta_g;\mathcal{D}_{1:N,1:K}}} \\ &\le \mathbb{E}_{\mathcal{D}_{\le K}}[B_{i,K}] + \lambda_{\text{inter}}\mathcal{I}^{\text{inter}}_{\theta_i \leftrightarrow \theta_g;\mathcal{D}_{1:N,1:K}} + \frac{\mathbb{E}_{\mathcal{D}_{\le K}}\left[\sigma^2_{H,i,K}(\mathcal{D}_{\le K})\right]}{2\lambda_{\text{inter}}}.\end{aligned}$$

This relates the existing $\lambda_{\text{inter}}\mathcal{I}^{\text{inter}}_{\theta_i \leftrightarrow \theta_g;\mathcal{D}_{1:N,1:K}}$ term in Eqn. (25) to the contribution of posterior dependence to the population bound. Finite mean concentration alone does not make the last term an optimization constant: the mean variance proxy may change with the candidate posterior family. Treating that term as uniformly bounded during optimization requires a common upper bound across the candidates. The expected baseline must likewise be retained. Appendix G.2 discusses these limitations.

B Additional Experiments

B.1 Ablation Study

We report results of ablation study in Table 2.

We additionally examine the roles of personalization and aggregation on *CIFAR-100* in Table 3. The variant without personalization relies on the global model, whereas the variant without aggregation performs purely local personalization. Bayes-pFCL achieves the highest accuracy among these three configurations. Removing aggregation yields lower forgetting but substantially lower accuracy, indicating that lower forgetting alone does not imply better overall predictive performance in this setting.

B.2 Sensitivity Tests

We report results of sensitivity tests in Table 4.

We further examine λ_{intra} and λ_{inter} at values 0.3, 1, and 10 on *CIFAR-100* and *TinyImageNet* under Dirichlet $\alpha = 1$; see Table 5. Across these tested values, accuracy decreases as either weight increases on both datasets. For λ_{inter}, forgetting also increases with the weight. For λ_{intra}, the lowest forgetting occurs at 1 among the tested values on both datasets, whereas increasing the weight to 10 reduces accuracy and increases forgetting.

Table 4: Sensitivity analysis of Bayes-pFCL with respect to the interference regularization weights on *EMNIST-Letters*, *CIFAR-100*, *TinyImageNet*, and *TCGA-BRCA*. We report *Accuracy* (%) and *Forgetting* (%) under different values of λ_{intra} and λ_{inter}.

Method	EMNIST-Letters		CIFAR-100		TinyImageNet		TCGA-BRCA	
	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓
$\lambda_{\text{inter}} = 0.5$	78.12 ± 3.05	3.58 ± 1.18	67.48 ± 2.43	2.51 ± 1.49	68.94 ± 3.66	4.31 ± 1.95	72.63 ± 2.48	3.16 ± 1.36
$\lambda_{\text{inter}} = 0.1$	80.41 ± 3.11	3.39 ± 1.23	69.83 ± 2.49	2.36 ± 1.51	71.24 ± 3.74	4.12 ± 1.99	74.56 ± 2.55	2.97 ± 1.38
$\lambda_{\text{intra}} = 0.5$	78.76 ± 3.08	3.21 ± 1.15	68.37 ± 2.47	2.18 ± 1.42	69.85 ± 3.70	3.97 ± 1.90	73.41 ± 2.53	2.81 ± 1.31
$\lambda_{\text{intra}} = 0.1$	80.02 ± 3.14	3.54 ± 1.28	69.57 ± 2.53	2.49 ± 1.58	71.03 ± 3.79	4.28 ± 2.07	74.21 ± 2.58	3.08 ± 1.44

Table 5: Extended sensitivity analysis of Bayes-pFCL on *CIFAR-100* and *TinyImageNet* under Dirichlet $\alpha = 1$. We report *Accuracy* (%) and *Forgetting* (%) for λ_{inter} and λ_{intra} at values 0.3, 1, and 10.

Weight setting	CIFAR-100		TinyImageNet	
	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓
$\lambda_{\text{inter}} = 0.3$	68.72 ± 2.46	2.43 ± 1.50	70.37 ± 3.70	4.20 ± 1.97
$\lambda_{\text{inter}} = 1$	66.31 ± 2.38	2.76 ± 1.45	67.74 ± 3.57	4.58 ± 1.90
$\lambda_{\text{inter}} = 10$	61.84 ± 3.26	3.41 ± 2.37	62.28 ± 3.35	5.36 ± 1.76
$\lambda_{\text{intra}} = 0.3$	68.91 ± 2.50	2.31 ± 1.47	70.31 ± 3.75	4.09 ± 2.01
$\lambda_{\text{intra}} = 1$	66.87 ± 2.41	1.96 ± 1.38	68.11 ± 3.63	3.69 ± 1.84
$\lambda_{\text{intra}} = 10$	62.35 ± 3.29	3.29 ± 2.20	62.91 ± 3.40	5.88 ± 2.59

B.3 Additional Baseline Comparisons

Evaluation setup. We extend the comparisons to FedAvg with client-side EWC [23], pFedDIL [31], and FedAGC [59]. Public source code for pFedDIL and FedAGC was unavailable at the time of our experiments; all results reported for these two methods in Table 6 are obtained from our own reimplementations based on their published descriptions. For the pFCL comparisons in Table 6, the numbers of clients, tasks per client, and classes per task, denoted by (N, K, C), are $(8, 6, 4)$ for *EMNIST-Letters*, $(10, 4, 20)$ for *CIFAR-100*, and $(10, 6, 30)$ for *TinyImageNet*. Tasks within each client have disjoint class sets, and we use the accuracy and forgetting metrics defined in the main experiments. The image experiments use ResNet-18 client models; for low-resolution inputs such as CIFAR-100, the input stem is adapted and the first max-pooling layer is removed. FedAvg + EWC applies EWC locally at each client to retain previous-task knowledge; its regularization coefficient and Fisher-information estimation follow the original EWC paper [23]. Each additional baseline experiment is repeated five times. These additional comparisons cover the three image benchmarks; corresponding TCGA-BRCA results are not available.

Results. Table 6 reports the additional baselines alongside the static and dynamic variants of Bayes-pFCL. Both variants achieve higher accuracy than the added baselines on all three datasets.

B.4 Effect of Client Heterogeneity

We examine the sensitivity of Bayes-pFCL to client heterogeneity by varying the Dirichlet concentration parameter α on *CIFAR-100* and *TinyImageNet*. Table 7 reports accuracy and forgetting for $\alpha \in \{0.1, 0.5, 1\}$, where smaller α corresponds to stronger client heterogeneity. Across these settings, accuracy increases and forgetting decreases as α increases on both datasets. These results characterize the sensitivity of Bayes-pFCL to heterogeneity within the tested range.

C Algorithms

Algorithm 1 gives the pseudocode of dynamic Bayes-pFCL for interference-regularized training with the objective in Eqn. (25), using full-model aggregation on a common parameter space. We write $\Theta_k^{(t)} = (\theta_{g,k}^{(t)}, \theta_{1,k}^{(t)}, \ldots, \theta_{N,k}^{(t)})$ and let $\mathcal{S}_k^{(t)}$ collect the distributed posterior summaries, including the global and local means and the precision blocks $\Lambda_{gg,k}^{(t)}$, $\Lambda_{ii,k}^{(t)}$, and $\Lambda_{gi,k}^{(t)}$. The task-start summaries

Table 6: Additional baseline comparisons under our pFCL task protocol. We report *Accuracy* (%) and *Forgetting* (%) on the three image datasets. *Reproduced* denotes our reimplementation. The Bayes-pFCL reference results are repeated from Table 1.

Method	**EMNIST-Letters**		**CIFAR-100**		**TinyImageNet**	
	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓
FedAvg + EWC	58.72 ± 4.09	5.80 ± 3.25	40.54 ± 4.83	5.31 ± 4.40	47.86 ± 2.59	4.59 ± 2.34
pFedDIL (Reproduced)	76.32 ± 3.32	4.73 ± 1.76	58.41 ± 2.67	5.01 ± 1.70	61.47 ± 4.58	4.38 ± 2.06
FedAGC (Reproduced)	72.61 ± 3.06	5.45 ± 1.84	45.47 ± 2.50	7.92 ± 1.85	44.69 ± 5.02	8.56 ± 2.77
Bayes-pFCL (Static)	79.34 ± 3.17	3.46 ± 1.26	68.92 ± 2.55	2.42 ± 1.56	70.17 ± 3.81	4.20 ± 2.03
Bayes-pFCL (Dynamic)	83.01 ± 3.24	3.34 ± 1.21	72.51 ± 2.28	2.24 ± 1.45	73.09 ± 4.39	5.57 ± 1.88

Table 7: Performance of Bayes-pFCL under different Dirichlet concentration parameters on *CIFAR-100* and *TinyImageNet*. We report *Accuracy* (%) and *Forgetting* (%) for $\alpha \in \{0.1, 0.5, 1\}$.

Dirichlet α	**CIFAR-100**		**TinyImageNet**	
	Accuracy (%)↑	Forgetting (%)↓	Accuracy (%)↑	Forgetting (%)↓
0.1	67.84 ± 2.91	3.72 ± 1.08	68.92 ± 4.21	5.96 ± 2.08
0.5	69.96 ± 2.47	2.78 ± 0.91	71.45 ± 3.78	4.72 ± 1.84
1	71.37 ± 2.22	2.15 ± 0.76	73.63 ± 3.46	3.84 ± 1.70

$\mathcal{S}_k^{(0)}$ are inherited from the preceding task and retained as a fixed reference throughout task k; for $k = 1$ they are initialized from the prior.

The penalty evaluations, fixed-reference convention, and gradient dependencies are specified in Appendix F.4. The interpretation of the online precision update is given in Appendix F.1.

D Compatibility of Bayes-pFCL with FCL and pFL Scenarios

D.1 Compatibility with Different pFL Categories.

Current pFL methods are mainly grouped into three categories: (1) regularization-based methods [30]; (2) personalized-head-based methods [60]; and (3) meta-learning-based methods [12]. Next, we show how to adapt Bayes-pFCL to these three categories.

Regularization-Based pFL. Regularization-based pFL adds a coupling regularizer $\mathcal{R}(\theta_g, \theta_i)$ to the empirical loss in Eqn. (25). Since such a regularizer only strengthens the dependency between θ_i and θ_g, without changing the model variables or their Bayesian dependency structure, Bayes-pFCL can be directly adapted to regularization-based pFL methods.

Personalized-Head-Based pFL. Personalized-head-based pFL usually decomposes each client model into a shared feature extractor and a personalized head, *i.e.*, $\theta_i = (\phi_i, \psi_i)$, where ϕ_i is aligned with the global model θ_g, while ψ_i remains client-specific. Accordingly, Bayes-pFCL can be adapted by applying the inter-model interference only to the shared component ϕ_i and θ_g, while applying the intra-model interference to the whole personalized model θ_i. This matches the objective of personalized-head-based pFL: the shared extractor captures cross-client knowledge, whereas the personalized head preserves client-specific knowledge. To further model the dependence between θ_g and θ_i accurately, one direction is to approximate the coupling relationship between ϕ_i and ψ_i based on the method-specific ways of personalizing heads. Here, we leave this as an open question, since the approximation details are method-specific and cannot be fully specified in a general form.

Meta-Learning-Based pFL. Meta-learning-based pFL treats the global model θ_g as an initialization or meta-prior from which each personalized model θ_i is adapted using local data. In the Bayesian view, this corresponds to conditioning each client posterior on the global initialization, *i.e.*, $p(\theta_i \mid \theta_g, \mathcal{D}_{i,1:K})$. Therefore, Bayes-pFCL can be adapted by measuring how the meta-initialization θ_g interferes with the client posterior after local adaptation, and how the new task $\mathcal{D}_{i,K}$ interferes with the posterior induced by previous tasks. Thus, Bayes-pFCL remains compatible with meta-learning-based pFL by regularizing the same intra-model and inter-model interference terms under the meta-initialization-based dependency.

Algorithm 1 Bayesian Personalized Federated Continual Learning (Bayes-pFCL)

Require: Clients $\{C_i\}_{i=1}^N$ with task sequences $\{\mathcal{D}_{i,k}\}_{k=1}^K$; communication rounds $\{\mathcal{T}_k\}_{k=1}^K$; local epochs E; learning rate η; weights $\rho_i \geq 0$, $\sum_i \rho_i = 1$; regularization weights $\lambda_{\text{intra}}, \lambda_{\text{inter}}$.

Ensure: Personalized models $\{\theta_i\}_{i=1}^N$ and global model θ_g.

Initialize $\Theta_1^{(0)} = (\theta_{g,1}^{(0)}, \theta_{1,1}^{(0)}, \ldots, \theta_{N,1}^{(0)})$ and prior summaries $\mathcal{S}_1^{(0)}$.
for task $k = 1, \ldots, K$ **do**
  if $k > 1$ **then**
    $\Theta_k^{(0)} \leftarrow \Theta_{k-1}^{(\mathcal{T}_{k-1})}$; $\mathcal{S}_k^{(0)} \leftarrow \mathcal{S}_{k-1}^{(\mathcal{T}_{k-1})}$.
  end if
  Preserve $\mathcal{S}_k^{(0)}$ as the fixed pre-task posterior reference throughout task k.
  for communication round $t = 1, \ldots, \mathcal{T}_k$ **do**
    Server broadcasts $\theta_{g,k}^{(t-1)}$ and synchronizes the required posterior context from $\mathcal{S}_k^{(t-1)}$, including $\Lambda_{gg,k}^{(t-1)}$.
    for each client $i = 1, \ldots, N$ in parallel **do**
      Initialize $\theta_{i,k}^{(t,0)} \leftarrow \theta_{i,k}^{(t-1)}$.
      for local epoch $e = 1, \ldots, E$ **do**
        Evaluate the candidate-dependent interference penalties $\mathcal{I}_{i,k}^{\text{intra},(t,e)}(\vartheta)$ and $\mathcal{I}_{i,k}^{\text{inter},(t,e)}(\vartheta)$ for Eqns. (13)–(14), using the synchronized context and the fixed pre-task reference as described in Appendix F.4.
        Take a local gradient step with the interference penalties:

$$\theta_{i,k}^{(t,e)} \leftarrow \theta_{i,k}^{(t,e-1)} - \eta \nabla_\vartheta \Big[\mathcal{L}_i(\vartheta; \mathcal{D}_{i,k}) + \lambda_{\text{intra}} \mathcal{I}_{i,k}^{\text{intra},(t,e)}(\vartheta) + \lambda_{\text{inter}} \mathcal{I}_{i,k}^{\text{inter},(t,e)}(\vartheta) \Big]_{\vartheta = \theta_{i,k}^{(t,e-1)}}.$$

      end for
      Set $\theta_{i,k}^{(t)} \leftarrow \theta_{i,k}^{(t,E)}$ and send the full model $\theta_{i,k}^{(t)}$ to the server.
    end for
    Aggregate $\theta_{g,k}^{(t)} \leftarrow \sum_{i=1}^N \rho_i \theta_{i,k}^{(t)}$ and broadcast $\theta_{g,k}^{(t)}$.
    Refresh the distributed Laplace summaries $\mathcal{S}_k^{(t)}$ at $\Theta_k^{(t)}$, including the global and local means and the blocks $\Lambda_{gg,k}^{(t)}$, $\Lambda_{ii,k}^{(t)}$, and $\Lambda_{gi,k}^{(t)}$ for all i.
  end for
  Retain $\Theta_k^{(\mathcal{T}_k)}$ and $\mathcal{S}_k^{(\mathcal{T}_k)}$ for the next task.
end for
return $\{\theta_{i,K}^{(\mathcal{T}_K)}\}_{i=1}^N$ and $\theta_{g,K}^{(\mathcal{T}_K)}$.

D.2 Compatibility with FCL.

Changing scenarios from pFCL to FCL makes θ_g to be the focus, rather than $\theta_g, \{\theta_i\}_{i=1}^N$; see Figure 3. Overall, we have the following static Bayesian framework for FCL.

Static Bayesian Framework of FCL. FCL combines the sequential task structure of CL with the multi-client structure of FL. Suppose client i observes a sequence of tasks $\{\mathcal{D}_{i,k}\}_{k=1}^K$. The goal is to infer a shared global parameter θ_g from all client data across all tasks. Accordingly, after observing all K tasks of all clients, the posterior is

$$p(\theta_g \mid \mathcal{D}_{1:N,1:K}) \propto p(\theta_g) \prod_{k=1}^K \prod_{i=1}^N p(\mathcal{D}_{i,k} \mid \theta_g),$$

where $\mathcal{D}_{1:N,1:K} := \{\mathcal{D}_{i,k}\}_{i=1,k=1}^{N,K}$.

Its recursive form is

$$p(\theta_g \mid \mathcal{D}_{1:N,1:K}) \propto p(\theta_g \mid \mathcal{D}_{1:N,1:K-1}) p(\mathcal{D}_{1:N,K} \mid \theta_g).$$

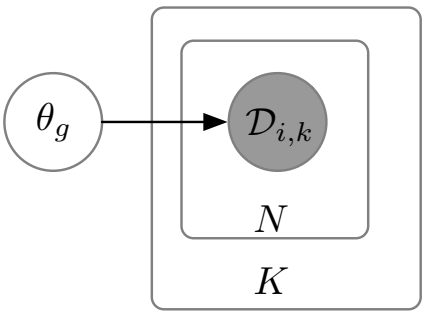


Figure 3: Bayesian PGM of FCL, where the model parameter θ_g accumulates evidence over sequential task datasets $\{\{\mathcal{D}_{i,k}\}_{i=1}^{N}\}_{k=1}^{K}$.

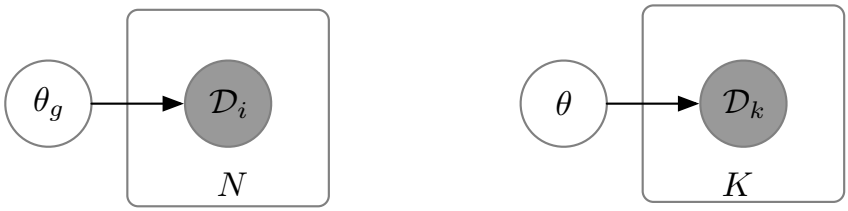


Figure 4: **Left:** Bayesian PGM of FL, where the global model θ_g aggregates evidence from distributed client datasets $\{\mathcal{D}_i\}_{i=1}^{N}$. **Right:** Bayesian PGM of CL, where the model parameter θ accumulates evidence over sequential task datasets $\{\mathcal{D}_k\}_{k=1}^{K}$.

This decomposition shows that the posterior induced by the first $k-1$ tasks across all clients serves as the prior for incorporating the current task k. Therefore, from a static Bayesian perspective, FCL can be viewed as recursively updating a global posterior while jointly accounting for continual adaptation and federated aggregation.

Via *Laplace* approximation, we can get the mode of the posterior as

$$\begin{aligned}\mu_K &= \arg\max_{\theta_g} \log p(\theta_g \mid \mathcal{D}_{1:N,1:K}) \\ &\approx \arg\max_{\theta_g} \left[\log p(\mathcal{D}_{1:N,K} \mid \theta_g) - \frac{1}{2}(\theta_g - \mu_{K-1})^\top \Lambda_{K-1}(\theta_g - \mu_{K-1})\right].\end{aligned}$$

E Bayesian Framework in FL and CL

Bayesian Frameworks primarily provide a probabilistic view of optimization. In CL, Bayesian inference provides a principled mechanism for sequential evidence accumulation, where the posterior after previous tasks serves as the prior for subsequent tasks [39, 24]. In FL, Bayesian inference provides a probabilistic mechanism for cross-client evidence aggregation, where client-side posterior information is combined to form global or personalized beliefs [58, 61, 34]. Thus, prior studies mainly use Bayesian tools to formalize learning dynamics and parameter estimation, but rarely use them to define knowledge and interference mathematically. Formally, given tasks $\{\mathcal{D}_k\}_{k=1}^{K}$ and parameter θ, CL can be described by the joint evidence $p(\mathcal{D}_{1:K} \mid \theta)$, where $\mathcal{D}_{1:K} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$; see Figure 4. Similarly, given clients $\{\mathcal{D}_i\}_{i=1}^{N}$ and global parameter θ_g, FL can be described by $p(\mathcal{D}_{1:N} \mid \theta_g)$, where $\mathcal{D}_{1:N} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$; see Figure 4.

F Practical Surrogates and Online Laplace Updates

This section specifies the Gaussian constructions used to evaluate the regularizers in Eqn. (25). We first clarify the online precision recursion following Theorem 3.3, then give the static and dynamic penalties and their use in Algorithm 1. The objective and the theoretical interference definitions remain as stated in the main text; the computational penalties below are distinguished from exact population mutual information.

F.1 Online Laplace Updates with a Fixed Gaussian Reference

Curvature of the full-history posterior. Theorem 3.3 evaluates every Hessian contribution at the same current MAP μ_K. Bayes' rule gives

$$\Lambda_{gi,K} = \nabla_{\theta_g}\nabla_{\theta_i}^{\top}\left[-\log p(z \mid \mathcal{D}_{1:N,1:K-1})\right]\big|_{z=\mu_K} + H_{gi,K},$$
$$H_{gi,K} := \nabla_{\theta_g}\nabla_{\theta_i}^{\top}\left[-\log p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i)\right]\big|_{z=\mu_K}.$$

For the full-history Laplace approximation, the preceding precision $\Lambda_{gi,K-1}$ is instead evaluated at μ_{K-1}. The old-posterior curvature at μ_K generally differs from its value at μ_{K-1}. Thus, the same-point Hessian decomposition alone does not imply an additive recursion using the stored old precision.

Online Gaussian construction. In the following construction, μ_K and Λ_K denote the mode and curvature obtained after replacing the old posterior by its fixed Gaussian approximation. They need not coincide with the MAP and Laplace precision of the original full-history posterior. Retain

$$q_{K-1}(z) := \mathcal{N}(z; \mu_{K-1}, \Lambda_{K-1}^{-1}), \qquad \Lambda_{K-1} \succ 0,$$

and hold this reference fixed while learning task K. Assume the product of this reference and the current-task likelihood is normalizable, attains a mode, and has a twice continuously differentiable log density in a neighborhood of that mode. Define

$$\mu_K := \arg\max_z \left[\log q_{K-1}(z) + \sum_{i=1}^{N} \log p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i)\right].$$

The negative log of q_{K-1} is a quadratic with constant Hessian Λ_{K-1}. Consequently, the curvature at this mode is

$$\Lambda_K := \Lambda_{K-1} + \nabla_z^2\left[-\sum_{i=1}^{N} \log p(\mathcal{D}_{i,K} \mid \theta_g, \theta_i)\right]\Bigg|_{z=\mu_K},$$
$$\Lambda_{gi,K} = \Lambda_{gi,K-1} + H_{gi,K}.$$

Here $H_{gi,K}$ is evaluated at the mode of this online construction. When the resulting joint curvature is positive definite, store

$$q_K(z) := \mathcal{N}(z; \mu_K, \Lambda_K^{-1})$$

as the Gaussian reference for the next task. Positive definiteness is a condition on the joint precision; the cross block $\Lambda_{gi,K}$ need not be square or positive definite. If a different positive-definite curvature approximation is used, the resulting matrix defines that Gaussian surrogate and need not equal the Hessian of the original posterior.

The online recursion in the main text is interpreted under this fixed-reference Gaussian construction. The preceding quadratic log-posterior expression is the log of this approximate update, up to normalization. It does not assert that old exact-posterior curvature is unchanged when the MAP moves. Within a task, a curvature refresh evaluates the current-task contribution against the fixed pre-task reference. The task-level recursion does not justify adding the curvature of the same task data repeatedly as if every communication round supplied an independent observation.

F.2 Static Gaussian Surrogates

For a fixed observed history, let q_{K-1} and q_K be the stored joint Gaussians. Their personalized marginals are

$$q_{i,k} = \mathcal{N}(\mu_{i,k}, \Lambda_{i,k}^{-1}), \qquad \Lambda_{i,k} := \left([\Lambda_k^{-1}]_{ii}\right)^{-1}, \qquad k \in \{K-1, K\}.$$

Thus, the marginal covariance is taken from the inverse of the full joint precision, rather than by inverting its diagonal block. The realized intra-model Gaussian penalty is

$$\widehat{\mathcal{I}}_{i,K}^{\mathrm{intra}} := \mathrm{KL}(q_{i,K} \| q_{i,K-1})$$
$$= \frac{1}{2}\Big[\mathrm{tr}(\Lambda_{i,K-1}\Lambda_{i,K}^{-1}) + \Delta\mu_{i,K}^{\top}\Lambda_{i,K-1}\Delta\mu_{i,K} - d_i + \log\frac{|\Lambda_{i,K}|}{|\Lambda_{i,K-1}|}\Big],$$

where $\Delta\mu_{i,K} := \mu_{i,K} - \mu_{i,K-1}$ and $d_i := \dim(\theta_i)$. For the inter-model penalty, write $\mathcal{D}_{\le K} := \mathcal{D}_{1:N,1:K}$. The population quantity in Appendix A.2 averages the conditional KL under a single joint law for the data and model parameters:

$$\mathcal{I}^{\text{inter}}_{\theta_i \leftrightarrow \theta_g; \mathcal{D}_{1:N,1:K}} = \mathbb{E}_{\mathcal{D}_{\le K}} \mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}} \left[\mathrm{KL}\Big(Q^{\theta_{-i}}_{i,K} \| P^{\theta_{-i}}_{i,K} \Big) \right].$$

For a fixed observed dataset, extract the principal block

$$\Lambda_{\{g,i\},K} := \begin{bmatrix} \Lambda_{gg,K} & \Lambda_{gi,K} \\ \Lambda_{ig,K} & \Lambda_{ii,K} \end{bmatrix}$$

from the same joint Gaussian precision. The conditional Gaussian and the product of its marginals, derived in Appendix F.5, give

$$\widehat{\mathcal{I}}^{\text{inter}}_{i,K}(\theta^*_{-i}; \mathcal{D}_{\le K}) = \frac{1}{2} \log \frac{|\Lambda_{gg,K}|\,|\Lambda_{ii,K}|}{|\Lambda_{\{g,i\},K}|}.$$

For this fixed joint Gaussian, the value does not depend on θ^*_{-i}. It is a fixed-data conditional information quantity; the precision blocks still depend on $\mathcal{D}_{\le K}$. When the conditional posterior family is Gaussian and the outer expectation uses the same data law, its population CMI is therefore

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}_{\le K}} \mathbb{E}_{\theta_{-i}|\mathcal{D}_{\le K}} \left[\widehat{\mathcal{I}}^{\text{inter}}_{i,K}(\theta_{-i}; \mathcal{D}_{\le K}) \right] \\ &\quad = \mathbb{E}_{\mathcal{D}_{\le K}} \left[\frac{1}{2} \log \frac{|\Lambda_{gg,K}|\,|\Lambda_{ii,K}|}{|\Lambda_{\{g,i\},K}|} \right]. \end{aligned}$$

This equality concerns the specified Gaussian conditional family and data law. If these Gaussians approximate the original posteriors, the averaged log-determinant remains a surrogate for the population CMI of the original posterior law. Neither a single observed log-determinant nor its Gaussian approximation is in general equal to that exact population quantity. Appendix A.2 states the KL guarantees for the chosen distributions and the additional Bayesian-consistency conditions needed for the intra-model MI interpretation.

F.3 Dynamic Gaussian Surrogates

We construct Gaussian penalties motivated by Eqns. (13) and (14). The task-level update in Appendix F.1 specifies an online approximation; it is not an exact filtering recursion for an arbitrary dynamic transition model. The round-specific Gaussian below also conditions on the observed parameter trajectory. Following the fixed-task notation in the main text, we omit K from the individual model parameters and write $\Theta^{(t)}_K := \big(\theta^{(t)}_g, \theta^{(t)}_1, \ldots, \theta^{(t)}_N\big)$. At communication round t of task K, assume

$$p\Big(\Theta^{(t)}_K \mid \Theta^{(0:t-1)}_K, \mathcal{D}_{1:N,1:K}\Big) \approx \mathcal{N}\Big(\Theta^{(t)}_K; \mu^{(t)}_K, (\Lambda^{(t)}_K)^{-1}\Big),$$

where

$$\mu^{(t)}_K := \big[(\mu^{(t)}_{g,K})^\top, (\mu^{(t)}_{1,K})^\top, \ldots, (\mu^{(t)}_{N,K})^\top\big]^\top.$$

The precision matrix follows the same parameter ordering:

$$\Lambda^{(t)}_K = \begin{bmatrix} \Lambda^{(t)}_{gg,K} & \Lambda^{(t)}_{g1,K} & \cdots & \Lambda^{(t)}_{gN,K} \\ \Lambda^{(t)}_{1g,K} & \Lambda^{(t)}_{11,K} & \cdots & \Lambda^{(t)}_{1N,K} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda^{(t)}_{Ng,K} & \Lambda^{(t)}_{N1,K} & \cdots & \Lambda^{(t)}_{NN,K} \end{bmatrix}.$$

Dynamic intra-model interference. For the fixed pre-task and round-t joint Gaussian approximations, define the personalized marginal precisions by

$$\Lambda^{(s)}_{i,K} := \Big(\Big[(\Lambda^{(s)}_K)^{-1} \Big]_{ii} \Big)^{-1}, \qquad s \in \{0, t\}.$$

The covariance block is taken from the inverse of the same joint precision matrix used for each Gaussian approximation. The corresponding personalized Gaussian marginals are

$$q^{(0)}_{i,K} := \mathcal{N}\Big(\mu^{(0)}_{i,K}, (\Lambda^{(0)}_{i,K})^{-1}\Big) \approx p(\theta^{(0)}_i \mid \mathcal{D}_{1:N,1:K-1}),$$

and

$$q_{i,K}^{(t)} := \mathcal{N}\Big(\mu_{i,K}^{(t)}, (\Lambda_{i,K}^{(t)})^{-1}\Big) \approx p(\theta_i^{(t)} \mid \Theta_K^{(0:t-1)}, \mathcal{D}_{1:N,1:K}).$$

Define the realized intra-model Gaussian KL surrogate by

$$\widehat{\mathcal{I}}_{i,K}^{\mathrm{intra},(t)} := \mathrm{KL}\Big(q_{i,K}^{(t)} \,\Big\|\, q_{i,K}^{(0)}\Big),$$

i.e.,

$$\widehat{\mathcal{I}}_{i,K}^{\mathrm{intra},(t)} = \frac{1}{2}\left[\mathrm{tr}\Big(\Lambda_{i,K}^{(0)}(\Lambda_{i,K}^{(t)})^{-1}\Big) + (\Delta\mu_{i,K}^{(t)})^\top \Lambda_{i,K}^{(0)} \Delta\mu_{i,K}^{(t)} - d_i + \log\frac{|\Lambda_{i,K}^{(t)}|}{|\Lambda_{i,K}^{(0)}|}\right],$$

where

$$\Delta\mu_{i,K}^{(t)} := \mu_{i,K}^{(t)} - \mu_{i,K}^{(0)}, \qquad d_i := \dim(\theta_i).$$

This quantity compares the round-t personalized marginal with the fixed pre-task marginal. If the old-task loss $f_{i,K}$ is sub-Gaussian under $q_{i,K}^{(0)}$, the old-reference KL bound in Appendix A applies to the corresponding Gaussian-surrogate expected loss increase. Identification with the population mutual information in Eqn. (13) additionally requires a coherent Bayesian model, averaging over task data and histories, marginalizing the trajectory, and a time-t prior predictive marginal given the pre-task data equal to $q_{i,K}^{(0)}$; the displayed conditional Gaussian KL does not automatically satisfy this identity.

Dynamic inter-model interference. Let $\theta_{-i}^{(t)} := \{\theta_j^{(t)}\}_{j\neq i}$. Conditioning on $\theta_{-i}^{(t)}$, the relevant precision block for $(\theta_g^{(t)}, \theta_i^{(t)})$ is

$$\Lambda_{\{g,i\},K}^{(t)} := \begin{bmatrix} \Lambda_{gg,K}^{(t)} & \Lambda_{gi,K}^{(t)} \\ \Lambda_{ig,K}^{(t)} & \Lambda_{ii,K}^{(t)} \end{bmatrix}.$$

For the fixed trajectory and data used to construct this joint Gaussian, define the inter-model surrogate by

$$\widehat{\mathcal{I}}_{i,K}^{\mathrm{inter},(t)} := \frac{1}{2}\left[\log|\Lambda_{gg,K}^{(t)}| + \log|\Lambda_{ii,K}^{(t)}| - \log|\Lambda_{\{g,i\},K}^{(t)}|\right].$$

This log-determinant is a fixed-data conditional information quantity of the chosen joint Gaussian, with the trajectory also held fixed. The population CMI in Eqn. (14) averages over the data and other personalized models, but does not condition on that trajectory. Its posterior must first marginalize the trajectory; the resulting mixture need not be Gaussian or have this same precision. Averaging the displayed log-determinant over data alone does not remove the trajectory conditioning. Even a joint data-and-trajectory average, under consistent conditioning laws, corresponds to an information quantity that still conditions on the trajectory, rather than the main-text CMI. The displayed log-determinant is therefore a training surrogate, without a general equality to the trajectory-marginalized population quantity.

F.4 Training-Time Evaluation and Optimization

Eqns. (13) and (14) motivate the training penalties. Algorithm 1 suppresses hats on its candidate-dependent interference penalties. Their numerical evaluations are the realized Gaussian surrogates defined below, with hats retained in this subsection to distinguish them from the exact population CMI. The local gradient step differentiates these Gaussian evaluations with respect to the candidate parameter; the notation does not assert evaluation of exact population CMI or its gradients. These penalties and their gradients are not claimed to be unbiased estimators of population CMI or its gradients. The task-start summaries $\mathcal{S}_k^{(0)}$ retain the fixed pre-task reference, and the inter-model penalty uses candidate same-round global and personalized variables. During local training, the round-t aggregate is not yet available. Clients therefore use the synchronized model and posterior context from round $t-1$ to construct differentiable Gaussian penalties at the current local iterate. This context is supplied through federated summaries; the conditioning on $\mathcal{D}_{1:N,1:k}$ does not require clients to access one another's raw data.

For a candidate personalized parameter ϑ, the Laplace-surrogate subroutine supplies a joint Gaussian with positive-definite precision. Its personalized marginal is denoted by $\widehat{q}_{i,k}^{(t,e)}(\vartheta)$, and its principal

precision block for (θ_g, θ_i) by $\widehat{\Lambda}^{(t,e)}_{\{g,i\},k}(\vartheta)$. The personalized marginal covariance is the ii block of the inverse of the same full joint precision; it is not the inverse of the candidate diagonal precision block. With the synchronized context suppressed in the notation, the closed forms in Appendix F.3 are evaluated as

$$\widehat{\mathcal{I}}^{\mathrm{intra},(t,e)}_{i,k}(\vartheta) = \mathrm{KL}\Big(\widehat{q}^{(t,e)}_{i,k}(\vartheta) \,\Big\|\, q^{(0)}_{i,k}\Big),$$

$$\widehat{\mathcal{I}}^{\mathrm{inter},(t,e)}_{i,k}(\vartheta) = \frac{1}{2}\log \frac{\left|\widehat{\Lambda}^{(t,e)}_{gg,k}(\vartheta)\right| \left|\widehat{\Lambda}^{(t,e)}_{ii,k}(\vartheta)\right|}{\left|\widehat{\Lambda}^{(t,e)}_{\{g,i\},k}(\vartheta)\right|}.$$

Here $q^{(0)}_{i,k}$ is the fixed pre-task posterior represented by $\mathcal{S}^{(0)}_k$, and the two diagonal precision blocks belong to the same candidate joint block. The local gradient is evaluated at $\vartheta = \theta^{(t,e-1)}_{i,k}$, retaining the candidate dependence of the current posterior and precision blocks, while holding the pre-task reference and synchronized context fixed. In particular, the inter-model gradient is taken through the candidate-dependent precision blocks rather than through a cached scalar. Posterior estimation and synchronization are written as subroutines; the curvature approximation is separate from this optimization and state-update specification, as discussed in Section 6.

F.5 *Laplace* Approximation in Assumption 3.7

We derive the Laplace form of

$$P^{\theta^*_{-i}}_{i,K} := p(\theta_g \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}) p(\theta_i \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}).$$

Applying the same joint *Laplace* construction as in Eqn. (8) at task K gives

$$p(z \mid \mathcal{D}_{1:N,1:K}) \approx \mathcal{N}\big(z; \mu_K, \Lambda_K^{-1}\big),$$

where

$$z := \big[\theta_g^\top, \theta_1^\top, \ldots, \theta_N^\top\big]^\top, \qquad \mu_K := \big[\mu_{g,K}^\top, \mu_{1,K}^\top, \ldots, \mu_{N,K}^\top\big]^\top.$$

Let

$$z_{gi} := \begin{bmatrix}\theta_g \\ \theta_i\end{bmatrix}, \qquad z_{-i} := \{\theta_j\}_{j\neq i}, \qquad \mu_{gi,K} := \begin{bmatrix}\mu_{g,K} \\ \mu_{i,K}\end{bmatrix}, \qquad \mu_{-i,K} := \{\mu_{j,K}\}_{j\neq i}.$$

Here, $z_{-i} = \theta_{-i}$ and $\mu_{-i,K}$ denote stacked column vectors, with clients $j \neq i$ ordered as in the full parameter vector z. Partition the joint precision matrix with respect to (z_{gi}, z_{-i}) as

$$\Lambda_K = \begin{bmatrix}\Lambda_{gi,gi,K} & \Lambda_{gi,-i,K} \\ \Lambda_{-i,gi,K} & \Lambda_{-i,-i,K}\end{bmatrix},$$

where

$$\Lambda_{gi,gi,K} = \begin{bmatrix}\Lambda_{gg,K} & \Lambda_{gi,K} \\ \Lambda_{ig,K} & \Lambda_{ii,K}\end{bmatrix}.$$

For a fixed θ^*_{-i}, the conditional posterior of (θ_g, θ_i) has the Gaussian approximation

$$Q^{\theta^*_{-i}}_{i,K} := p(\theta_g, \theta_i \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}) \approx \mathcal{N}\left(z_{gi}; \mu^{\theta^*_{-i}}_{gi,K}, \left(\Lambda^{\theta^*_{-i}}_{gi,K}\right)^{-1}\right),$$

with conditional precision

$$\Lambda^{\theta^*_{-i}}_{gi,K} = \Lambda_{gi,gi,K} = \begin{bmatrix}\Lambda_{gg,K} & \Lambda_{gi,K} \\ \Lambda_{ig,K} & \Lambda_{ii,K}\end{bmatrix},$$

and conditional mode

$$\mu^{\theta^*_{-i}}_{gi,K} = \mu_{gi,K} - \Lambda^{-1}_{gi,gi,K}\Lambda_{gi,-i,K}\left(\theta^*_{-i} - \mu_{-i,K}\right).$$

Equivalently, write

$$\mu^{\theta^*_{-i}}_{gi,K} = \begin{bmatrix}\mu^{\theta^*_{-i}}_{g,K} \\ \mu^{\theta^*_{-i}}_{i,K}\end{bmatrix}.$$

The decoupled posterior in Assumption 3.7 removes the conditional dependence between θ_g and θ_i by taking the product of the conditional marginals of this same $Q_{i,K}^{\theta^*_{-i}}$:

$$P_{i,K}^{\theta^*_{-i}} = p(\theta_g \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}) p(\theta_i \mid \theta^*_{-i}, \mathcal{D}_{1:N,1:K}).$$

Therefore,

$$P_{i,K}^{\theta^*_{-i}} \approx \mathcal{N}\left(\mu_{P_{i,K}}^{\theta^*_{-i}}, \left(\Lambda_{P_{i,K}}^{\theta^*_{-i}}\right)^{-1}\right),$$

where the decoupled mode is

$$\mu_{P_{i,K}}^{\theta^*_{-i}} := \begin{bmatrix} \mu_{g,K}^{\theta^*_{-i}} \\ \mu_{i,K}^{\theta^*_{-i}} \end{bmatrix},$$

and the decoupled precision is block diagonal:

$$\Lambda_{P_{i,K}}^{\theta^*_{-i}} := \begin{bmatrix} \Lambda_{g,K}^{\theta^*_{-i}} & 0 \\ 0 & \Lambda_{i,K}^{\theta^*_{-i}} \end{bmatrix}.$$

Here,

$$\left(\Lambda_{g,K}^{\theta^*_{-i}}\right)^{-1} = \left[\left(\Lambda_{gi,K}^{\theta^*_{-i}}\right)^{-1}\right]_{gg}, \qquad \left(\Lambda_{i,K}^{\theta^*_{-i}}\right)^{-1} = \left[\left(\Lambda_{gi,K}^{\theta^*_{-i}}\right)^{-1}\right]_{ii}.$$

Equivalently, by the Schur complement,

$$\Lambda_{g,K}^{\theta^*_{-i}} = \Lambda_{gg,K} - \Lambda_{gi,K}\Lambda_{ii,K}^{-1}\Lambda_{ig,K},$$

and

$$\Lambda_{i,K}^{\theta^*_{-i}} = \Lambda_{ii,K} - \Lambda_{ig,K}\Lambda_{gg,K}^{-1}\Lambda_{gi,K}.$$

Thus, $P_{i,K}^{\theta^*_{-i}}$ has mode $\mu_{P_{i,K}}^{\theta^*_{-i}}$ and precision $\Lambda_{P_{i,K}}^{\theta^*_{-i}}$, supplying the Gaussian construction for Assumption 3.7. These are also the reference mean and precision used for $B_{i,K}$ and Gaussian concentration at each fixed dataset in Theorem A.9; each Gaussian mean equals its mode. The population bound additionally averages the data-dependent concentration proxy $\sigma^2_{H,i,K}(\mathcal{D}_{\leq K})$ under the specified data law.

G Discussion

G.1 Potential Social Impact

Bayes-pFCL can benefit privacy-sensitive applications, such as healthcare and finance, by enabling clients to collaboratively learn personalized continual models without sharing raw data. This may improve model adaptation to heterogeneous local populations and evolving task distributions. Potential risks include privacy leakage from shared posterior summaries or model updates, amplification of client-specific biases under heterogeneous data, reliability issues caused by approximate posterior estimation, and increased computational cost. Thus, deployment of Bayes-pFCL in high-stakes domains should be accompanied by privacy-preserving mechanisms, fairness auditing, calibration checks, and domain-specific validation.

G.2 Scope of the Regularization Guarantees

Theorem 3.8 motivates the intra-model penalty through an expected old-task loss bound and the inter-model penalty through a bound on $\mathbb{E}_{\mathcal{D}_{\leq K}}[|H_{i,K} - B_{i,K}|]$. Here $H_{i,K}$ and $B_{i,K}$ are evaluated for each fixed dataset, while the inter-model guarantee averages over the same data law as the population CMI. Its concentration term is $\mathbb{E}_{\mathcal{D}_{\leq K}}[\sigma^2_{H,i,K}(\mathcal{D}_{\leq K})]$, which must be finite; the concentration proxy may vary across datasets. Positive-definite Gaussian precision alone does not establish the required loss concentration for unrestricted deep-network losses or exact posteriors.

The population baseline $\mathbb{E}_{\mathcal{D}_{\leq K}}[B_{i,K}]$ may vary with the conditional marginals, so reducing interference alone need not reduce $\mathbb{E}_{\mathcal{D}_{\leq K}}[H_{i,K}]$. Both the averaged baseline and the averaged concentration proxy can also depend on the candidate posterior family. They cannot be discarded as optimization

constants without establishing the corresponding invariance or an appropriate bound uniform over the candidates. Eqn. (25) combines empirical task fitting with two theoretically motivated interference penalties; the bounds do not establish equivalence to minimizing the full heterogeneity bound or guarantee monotone improvements in forgetting or heterogeneity during training. Appendix A.2 gives the concentration conditions, the KL bound, fixed-history information-gain bound, and population MI bounds, together with the Young inequalities relating them to the linear penalties. Appendix F distinguishes the observed-data Gaussian penalties from the population quantities.

H Hyperparameters

Table 8 reports the dataset configurations and training hyperparameters, including the number of clients N, tasks per client K, and classes per task C.

Table 8: Dataset configurations and training hyperparameters of Bayes-pFCL. "General" gives the default value when no dataset-specific value is specified. The total communication budget is $\mathcal{T} = \sum_{k=1}^{K} \mathcal{T}_k$, where $\mathcal{T}_k$ is the number of rounds for task k in Algorithm 1.

Description		**General**	**EMNIST-Letters**	**CIFAR-100**	**TinyImageNet**	**TCGA-BRCA**
N	# of clients	-	8	10	10	10
K	Tasks per client	-	6	4	6	4
C	Classes per task	-	4	20	30	2
$\mathcal{T}$	Total rounds	-	400	400	100	100
E	Local epochs	1	-	-	2	-
B	Local batch size	64	-	-	-	-
η	Local learning rate	1e-4	-	-	-	-
λ_{inter}	Balance term	1	-	-	-	-
λ_{intra}	Balance term	1	-	-	-	-

I Implementation

All the experiments are conducted under two RTX 3090 GPUs in Docker environments.

All methods are implemented in PyTorch [40] using the same neural network architecture for components with identical roles. Specifically, we use CNN [27] as feature extractor and ResNet18 [16] as classifier. Method-specific hyper-parameters are set following the original configurations of the respective baselines.

J Dataset Details

EMNIST-Letters [8] is a handwritten character recognition dataset with 26 letter classes from "A" to "Z." Each class contains approximately $4{,}800$ training samples and 800 test samples, and each sample is a 28×28 grayscale image.

CIFAR-100 [25] is a fine-grained natural-image classification dataset with 100 fine classes grouped into 20 coarse superclasses. Each fine class contains 500 training images and 100 test images, and each sample is a 32×32 RGB image. We use the fine labels for training and evaluation.

TinyImageNet [26] is a large-scale image classification dataset with 200 classes derived from the WordNet hierarchy. Each class contains 500 training samples and 50 validation samples, and each sample is a 64×64 RGB image.

TCGA-BRCA [49] is a real-world breast invasive carcinoma dataset from The Cancer Genome Atlas. It contains $1{,}095$ samples, where each sample includes over $18{,}000$ RNA-Seq gene-expression features, 39 clinical features, binary subtype labels for Basal-like versus non-Basal tumors, and demographic or clinical covariates such as age.

K Reproduction

The code is available at `https://github.com/IntelliSys-Lab/Bayes-pFCL`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The core contribution of our work is to mathematically frame knowledge and interference in pFCL as probabilistic models and develop a general method based on that to mitigate catastrophic forgetting and data heterogeneity.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the computational limitations of Bayes-pFCL, and Appendix G.2 discusses the scope of its regularization guarantees.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix A provides comprehensive theoreical analysis.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The supplementary materials include all the necessary codes. Also, the algorithm in Appendix C describes the whole workflow.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The supplementary materials include all necessary codes.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix H provides the configuration of hyperparameters

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results in Table 1 include both mean and std.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix I provides the information of detailed implementation and running environments of Bayes-pFCL.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Codes are fully anonymous.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G provides the discussion on potential social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All model and datasets used in our work are open-source and public available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing nor research with human subjects involved

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is only used for writing and formating purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.