EVIDENCE-BASED MODEL ALIGNMENT FOR UNCOORDINATED ONE-SHOT FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-Shot Federated Learning is a paradigm in which a global model is learned from data distributed across multiple clients in a single communication round. This setting becomes particularly challenging when the training process is uncoordinated, i.e., when clients start from different initializations and train for different numbers of epochs. We analyze this problem from a Bayesian perspective and show how accounting for the permutation symmetries of client posteriors improves the resulting global model. Based on these observations, we propose FLAME, a principled approach that identifies client permutations for model alignment by maximizing the Bayesian evidence of the merged model, and then performs Bayesian merging, thus handling permutation symmetries and weight uncertainty in a unified manner. We show on standard benchmarks that our proposal outperforms competing approaches that typically rely either on permutation-based matching or on client loss curvature alone.

1 INTRODUCTION

Federated Learning (FL) typically involves a server broadcasting a common model and aggregating client updates over multiple rounds of communication (McMahan et al., 2017; Kairouz & McMahan, 2021). The underlying problem is more general, and consists on effectively combining several models into a single global model that captures what they have individually learned. More specifically, models trained independently differ based on three dimensions: (i) the data they were trained on, a problem known as statistical heterogeneity and thoroughly tackled in FL-related literature (Kairouz & McMahan, 2021; Marfoq et al., 2022); (ii) the training procedure, e.g., initialization and duration, with such problem tackled by model alignment and merging literature through the permutation symmetries of neural networks (Ainsworth et al., 2023; Crisostomi et al., 2024); and (iii) the architecture itself, problem that we leave aside in this work (Rinaldi et al., 2026).

Standard FL typically handles the second dimension by construction: clients share an initialization and repeatedly receive the aggregated model over multiple rounds, keeping their local models in a common region of parameter space. This is a consequence of the protocol rather than an inherent property of the underlying problem, and it comes at a cost: repeated server-client communication is a major bottleneck in FL (Guha et al., 2019; Kairouz & McMahan, 2021). One-Shot FL (Guha et al., 2019) removes such a bottleneck by obtaining the global model in a single round, but in doing so becomes more sensitive to differences in clients' training that can place their models in different regions of parameters. This happens in practical scenarios where clients are *uncoordinated*: client models may have been trained before any federation existed or initialized from different pre-trained checkpoints, and heterogeneous devices may stop local training at different points (Nguyen et al., 2023; Yang et al., 2022). Clients then differ in both data and training, and a single round leaves no opportunity to reconcile these differences afterwards (Zeng et al., 2025). We study this setting, where the first two dimensions discussed above arise together and must be addressed jointly.

Existing methods tackle the challenge of model merging from different perspectives. Curvature-aware methods incorporate second-order information, i.e., the curvature of the local loss, which measures how precisely the local data determine each parameter when combining client models Jhunjhunwala et al. (2024); Talpini et al. (2025). However, differences in training procedures can leave each client's local description of the loss centered at a different point in parameter space, making direct combination unreliable and limiting the quality of the resulting model. Model align-

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

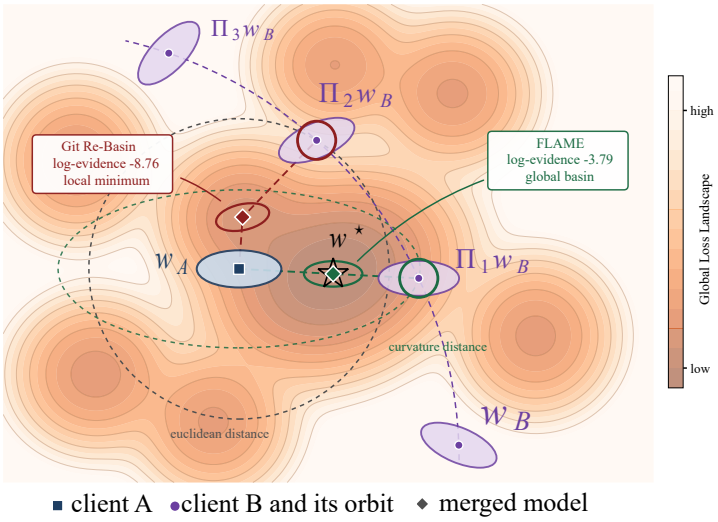

Figure 1: Depicting One-Shot FL with two clients trained from different initializations, each summarized by its local Laplace posterior (mean and curvature ellipse), over the "ground-truth" global loss. The functionally equivalent representations $\Pi\hat{w}_B$ of client B have equal norm and lie on the dashed arc; each merges with $\hat{w}_A$ into a different global model (diamonds). Euclidean distance selects $\Pi_2\hat{w}_B$ and merges into a local minimum, while the evidence of the merged model selects $\Pi_1\hat{w}_B$ and reaches the global basin at $w^\star$: the criterion that ranks alignments correctly is a property of the merge, not of the clients separately.

ment and merging methods address this parameter-space mismatch in depth Liu et al. (2022); Singh & Jaggi (2020): they exploit permutation symmetries to establish correspondences between client models before merging them Ainsworth et al. (2023); Crisostomi et al. (2024), but generally do not account for either the statistical heterogeneity nor the posterior uncertainty of the local solutions.

Therefore, neither parameter alignment nor local curvature alone provides a complete picture of the merging problem in One-Shot FL, especially in the challenging setting of uncoordinated training. We argue that model alignment and model merging should not be treated in isolation but as parts of a single inference problem, and frame both within a unified Bayesian perspective in which permutations are selected by maximizing the evidence of the merged model, starting from Laplace approximations of the clients' posteriors. Figure 1 illustrates this perspective: two local optima of Client B are related by a permutation and therefore represent the same local model, yet their corresponding posterior curvatures encode parameter uncertainty and can lead to different outcomes when merged with Client A. Parameter-space similarity cannot distinguish between these equivalent parameterizations, whereas Bayesian evidence evaluates their effect on the resulting merged model and favors the alignment that yields the better global solution.

Contributions. We summarize our main contributions as follows:

- We study One-Shot FL under *uncoordinated* training, where clients differ in initialization and number of local epochs, and cast alignment and aggregation as a single Bayesian inference problem.
- We propose FLAME (Federated Learning with Alignment by Model Evidence), which selects client permutations by maximizing the Bayesian evidence of the merged model and uses the same posteriors for the merge. We derive the evidence in closed form under Laplace approximations of the client posteriors and show that, under isotropic posteriors, it reduces to the weight-matching objective of common approaches Ainsworth et al. (2023); Crisostomi et al. (2024).
- We show that FLAME outperforms both alignment-based and curvature-aware baselines on standard benchmarks.

2 Related Work

FLAME sits at the intersection of One-Shot FL and model alignment and merging. Accordingly, we organize the related work in these areas by how each method combines models.

2.1 One-Shot Federated Learning

One-Shot FL methods fall into three families: knowledge distillation, probabilistic and neuron matching approaches. Methods from the first category distill the ensemble of client models into a global model using data at the server, either assumed available (Gong et al., 2021; Lin et al., 2020; Guha et al., 2019; Hasan et al., 2024) or synthesized by a generative model (Zhou et al., 2020; Heinbaugh et al., 2023). Since server-side data availability raises privacy concerns and/or may not represent the clients' distributions (Jhunjhunwala et al., 2024), we focus on data-free methods.

Probabilistic approaches aggregate clients through their local posteriors, making the curvature of the local loss explicit and using it to weight the parameters of the merged model (Liu et al., 2024; Jhunjhunwala et al., 2024; Talpini et al., 2025). In Jhunjhunwala et al. (2024) each client sends a local minimum together with a Fisher-based estimate of the curvature around it, and the server minimizes the sum of the resulting quadratic approximations by SGD, regularized towards the mean of the local parameters. Talpini et al. (2025) take a fully Bayesian approach, using an ensemble to capture multiple optima of both local and global loss and marginalizing over it at prediction time. All these methods assume that the local approximations describe comparable regions of parameter space, which may not hold when clients do not share an initialization.

Neuron matching approaches exploit the invariance of neural networks to permutations of their hidden units: client models are first aligned to a common ordering and then averaged Yurochkin et al. (2019); Singh & Jaggi (2020); Liu et al. (2022). Even if these solutions can handle independently trained clients, the alignment at the core of these methods is based on a matching criterion that measures overlap in parameter space, independently of how strongly each direction is supported by the local data, and the subsequent averaging is curvature-agnostic.

FLAME both leverages permutation symmetries and exploits local posteriors, selecting for each client the permutation that maximizes the evidence of the merged model.

2.2 Model Alignment and Merging

Model merging combines several trained networks in parameter space, without access to the training data and ideally without retraining. Methods differ in two perspective: how to align the networks and how to merge them.

Model alignment. Networks trained from different initializations may implement similar functions while lying in different regions of parameter space, because of the permutation symmetry described in Section 3.2; averaging their weights then combines unrelated units and destroys performance (Ainsworth et al., 2023). Studies of linear mode connectivity, i.e., whether the loss remains low along the straight line between two solutions, show that such networks can often be brought into a shared basin, i.e., a connected low-loss region, once symmetry is taken into account (Entezari et al., 2022; Ainsworth et al., 2023). The aligning permutation is estimated as an optimal transport problem between neurons (Singh & Jaggi, 2020), by weight- or activation-matching with a cycle-consistent extension to more than two models (Ainsworth et al., 2023; Crisostomi et al., 2024), or as a graph-matching problem (Liu et al., 2022). All these criteria measure overlap in the Euclidean geometry of parameter space, independently of how strongly each direction is supported by the data.

Model merging. Once networks are aligned, or when they share a pre-trained initialization and thus already lie in a common basin, the question becomes how to combine their parameters. Beyond plain averaging, existing methods correct the variance mismatch it introduces (Jordan et al., 2023), refine it by layer-wise least squares (Nasery et al., 2025), merge features beyond one-to-one permutations (Stoica et al., 2024), or resolve sign and magnitude conflicts between task vectors (Yadav et al., 2023). Closest to our approach, Fisher merging (Matena & Raffel, 2022) and RegMean (Jin et al., 2023) weight parameters by second-order statistics, but assume a shared basin and do not address misalignment.

FLAME makes both decisions with a single criterion: it aligns models in the metric induced by the clients' curvature rather than the Euclidean one, and combines them with the same curvature through Bayesian merging.

3 BACKGROUND

One-Shot FL considers a set of K clients that collaboratively learn a common predictive model without sharing their local data. Each client c has access to a private dataset $\mathcal{D}_k = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_k}$ and independently trains a model by minimizing a task-specific loss, commonly the negative log-likelihood associated with the assumed probabilistic model $p(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$ Murphy (2023); Al-Shedivat et al. (2021). The learned parameters $\hat{\mathbf{w}}_k$ are then communicated *once* to a central server, which combines them into a single global model. A principled way to perform this combination is through Bayesian inference. Given K datasets and the local posteriors $p(\mathbf{w}|\mathcal{D}_k)$ of the same model, under the assumptions that (i) the datasets are conditionally independent given the model and (ii) the prior is shared across clients, the global posterior can be written as Talpini et al. (2025);

$$p(\mathbf{w}|\mathcal{D}_1, .., \mathcal{D}_K) = \alpha \cdot \frac{\prod_{k=1}^{K} p(\mathbf{w}|\mathcal{D}_k)}{p(\mathbf{w})^{K-1}}, \tag{1}$$

which exactly recovers the posterior that would be obtained from the union of the clients' data. We refer to this operation as *Bayesian merging*.

3.1 LAPLACE APPROXIMATION AND MODEL EVIDENCE

To make Equation (1) tractable, it is common to approximate the posterior with a Gaussian, via the Laplace approximation Ritter et al. (2018): expanding the local loss to second order around a minimum $\hat{\mathbf{w}}_k$ yields $p(\mathbf{w}|\mathcal{D}_k) \approx \mathcal{N}(\hat{\mathbf{w}}_k, \Sigma_k)$, where the precision $\Lambda_k = \Sigma_k^{-1}$ is the Hessian of the local negative log-posterior at $\hat{\mathbf{w}}_k$, i.e., the curvature of the local loss. Beyond a tractable posterior, this gives access to the *model evidence*. For a model M, this is $p(\mathcal{D}|M) = \int p(\mathcal{D}|\mathbf{w}, M)\, p(\mathbf{w}|M)\, d\mathbf{w}$, i.e., the quantity we use to drive the alignment (see Section 4). Under the Laplace approximation centered at a local maximum $\hat{\mathbf{w}}$, its logarithm can be estimated, up to additive constants, as:

$$\ln p(\mathcal{D}|M) = \ln p(\mathcal{D}|\hat{\mathbf{w}}) + \ln p(\hat{\mathbf{w}}) - \tfrac{1}{2} \ln |\Lambda|. \tag{2}$$

Equation (2) has a well-known reading MacKay (2003); Murphy (2023): among models attaining a similar data fit (first term), it favours those whose posterior is broader (last term).

Finally, the product of Gaussians is itself an (unnormalized) Gaussian: for two local posteriors $\mathcal{N}(\mathbf{a}, A)$ and $\mathcal{N}(\mathbf{b}, B)$, the global posterior (Equation (1)) depends on their product $\mathcal{N}(\mathbf{c}, C) \propto \mathcal{N}(\mathbf{a}, A)\mathcal{N}(\mathbf{b}, B)$, with:

$$C = (A^{-1} + B^{-1})^{-1}, \qquad \mathbf{c} = C\,(A^{-1}\mathbf{a} + B^{-1}\mathbf{b}). \tag{3}$$

Dividing by the prior keeps the global posterior Gaussian, $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\Sigma^{-1} = \sum_k \Sigma_k^{-1} - (K-1)\,\sigma_p^{-2} I$ and $\boldsymbol{\mu} = \Sigma \sum_k \Sigma_k^{-1} \hat{\mathbf{w}}_k$; its mode $\boldsymbol{\mu}$ is the global model, and its normalizing constant gives the evidence of the *merged* model in closed form (see Section 4.1). We remark that model evidence is a well-established metric adopted for hyperparameter optimization MacKay (2003); Murphy (2023), adopted in applications such as Gaussian process kernel learning and in settings where conventional cross-validation is unavailable or impractical Immer et al. (2021).

3.2 PERMUTATION SYMMETRY

Neural networks are typically built upon layered architectures of matrix-vector products followed by simple nonlinearities, a structure that gives rise to symmetries in the parameter space: permuting the rows of a parameter matrix and applying the same permutation to the columns of the next one leaves the function computed by the network unchanged Ainsworth et al. (2023), and the posterior over parameters inherits this symmetry Bernacchia (2025). We denote by $\mathcal{P}$ the group of such function-preserving permutations Π, which is discrete and contains $d!$ elements for a ReLU layer with d neurons Hecht-Nielsen (1990). A local posterior therefore admits a whole family of symmetry-related Gaussian representations

$$\left\{\mathcal{N}\left(\Pi\hat{\mathbf{w}}_k,\, \Pi\Sigma_k\Pi^\top\right) : \Pi \in \mathcal{P}\right\}. \tag{4}$$

Although functionally equivalent, they are generally distinct as distributions in the parameter space. Consequently, directly combining the local posteriors in Equation (1) can lead to a poor merge if symmetry-equivalent parameters are not properly handled. The permutation symmetry must therefore be explicitly accounted for when combining local posteriors: rather than arbitrarily selecting one representative from each family, we seek the permutation that best aligns the corresponding posterior representations before performing the Bayesian merge. This alignment constitutes the central challenge addressed by our approach, as detailed in the next Section.

4 OUR PROPOSAL: FLAME

4.1 EVIDENCE-BASED SELECTION OF THE PERMUTATIONS

For simplicity, let us consider two clients that approximate their posteriors via the Laplace approximation as $\mathcal{N}(\mathbf{a}, A)$ and $\mathcal{N}(\mathbf{b}, B)$. As discussed in Section 3, their product is Gaussian, and we can compute the log-evidence explicitly. Without loss of generality (see Proposition A.5), we can keep $\mathcal{N}(\mathbf{a}, A)$ fixed, i.e., considering the identity as its permutation, and look for the permutation acting on $\mathcal{N}(\mathbf{b}, B)$ that maximizes the log-evidence, Z, of the merged model. More specifically, the following holds.

Proposition 4.1 (Log-evidence of the merged model)**.** *Let $\mathbf{c}_\Pi$ and C_Π be the mean and covariance of the product of clients' posterior, given by Equation (3) applied to $(\mathbf{a}, A)$ and $(\Pi\mathbf{b}, \Pi B\Pi^\top)$, i.e., the mean and covariance of the normalized product of the two local posteriors, and let the prior be $\mathcal{N}(\mathbf{0}, \sigma_p^2 I)$, shared across clients. Dropping terms constant in Π, the log-evidence is:*

$$Z(\Pi) \;=\; \underbrace{-\tfrac{1}{2}(\mathbf{a}-\Pi\mathbf{b})^\top (A+\Pi B\Pi^\top)^{-1}(\mathbf{a}-\Pi\mathbf{b})}_{\textit{Misalignment Term}} \;\; \underbrace{-\tfrac{1}{2}\ln\left|A+\Pi B\Pi^\top\right| - \tfrac{1}{2}\ln\left|\sigma_p^2 I - C_\Pi\right|}_{\textit{Volume/Occam Term}} + \\ + \underbrace{\tfrac{1}{2}\mathbf{c}_\Pi^\top(\sigma_p^2 I - C_\Pi)^{-1}\mathbf{c}_\Pi}_{\textit{Prior-location Term}}. \tag{5}$$

The proof is given in Appendix A and follows from Theorem A.2 in Appendix A.2. The three terms have a direct reading. The *misalignment* term penalizes discrepancies between clients means, weighted by their combined uncertainty, so that disagreements along low-uncertainty directions are penalized more strongly. The two log-determinants of the *Volume/Occam* term together equal, up to a constant, the log-volume $\frac{1}{2}\ln|\Sigma_\Pi|$ of the global posterior (Lemma A.4), favoring alignments that yield a broad merged posterior. Finally, the *prior-location* term grows with the norm of c_Π, favoring permutations for which the aligned client parameter vectors have similar directions.

Accordingly, FLAME selects

$$\Pi^\star \;=\; \arg\max_{\Pi\in\mathcal{P}} \; Z(\Pi). \tag{6}$$

4.1.1 A GEOMETRIC VIEW: ALIGNMENT AS FISHER-METRIC WEIGHT MATCHING

We have derived FLAME probabilistically; in this Section we discuss Equation (5) as an objective function from a geometric perspective, showing that under some assumptions, optimizing for Equation (5) recovers well-known approaches of matching literature Ainsworth et al. (2023).

By permutation symmetry, each local posterior represents an orbit of functionally equivalent Gaussians (Equation (4)). Aligning two clients means choosing the representatives that overlap best, and the notion of "overlap" is set by a metric. This is what separates classical weight matching from FLAME: Git Re-Basin Ainsworth et al. (2023) measures overlap with the flat Euclidean metric, whereas FLAME uses the curvature-induced metric of the merge. We further analyze the behavior of the evidence in the broad-prior limit, i.e., as $\sigma_p^2 \to \infty$: Lemma A.6 shows that the prior-location term and the second contribution to the volume/Occam term in Equation (5) become permutation-invariant. Thus, in this limit, maximizing Z is equivalent to minimizing the misalignment term computed as a squared Mahalanobis distance, while maximizing the volume of the merged posterior. In addition, assuming an isotropic posterior, we can recover Ainsworth et al. (2023), as specified by:

Proposition 4.2 (Weight matching as isotropic case)**.** *Let the local posteriors be isotropic, $A = \sigma_A^2 I$ and $B = \sigma_B^2 I$, and let the prior be $\mathcal{N}(0, \sigma_p^2 I)$. Then $Z(\Pi) = \beta\langle a, \Pi b\rangle + \text{const}$ with $\beta > 0$, hence*

$$\arg\max_{\Pi\in\mathcal{P}} Z(\Pi) = \arg\min_{\Pi\in\mathcal{P}} \|a - \Pi b\|_2^2 = \arg\max_{\Pi\in\mathcal{P}} \langle a, \Pi b\rangle, \tag{7}$$

That is, evidence maximization reduces to the weight-matching objective of Git Re-Basin (Ainsworth et al., 2023).

4.2 THE FLAME ALGORITHM

FLAME operates in three stages: (i) the server optimizes the client permutations by maximizing the evidence of the merged model (Equation (5)); (ii) it merges the aligned Laplace posteriors according to Equation (1); and (iii) it returns the mode of the resulting global posterior to the clients as the global model. Every stage uses only the clients' Laplace approximations, so the whole procedure is data-free at the server. Algorithm 1, reported in Appendix B.1, details the procedure; below, we describe each stage.

Permutation optimization. The core operation of FLAME is the maximization of the merged evidence over the permutations of all clients. The optimization variables are one permutation matrix per client and per hidden layer, which reorders the neurons of that layer.

Diagonal evidence. Although the clients' Laplace approximations use Kronecker-factored curvature, we compute the evidence for alignment with its diagonal. Under this approximation, each parameter carries its own uncertainty, and every term of Equation (5) becomes a sum of per-parameter contributions. For a given layer, the cost of matching neuron i of one client to neuron j of another then depends on the weights attached to them, i.e., their incoming and outgoing connections and their corresponding uncertainty. Thus, with a diagonal approximation, aligning a layer reduces to finding the optimal one-to-one matching of neurons, which is a linear assignment problem. Despite this simplification, FLAME still outperforms Euclidean matching-based alignment (see Section 5.2).

Coupling and optimization. The linear-assignment structure holds for one layer at a time and only once the rest is fixed. First, a permutation of layer ℓ reorders both the outgoing weights of layer ℓ and the incoming weights of layer $\ell + 1$, so the best permutation of a layer depends on those of its neighbors. Second, with more than two clients, the evidence contribution of each client depends on the merged model, which itself depends on the permutations of all clients (Theorem A.2). We therefore follow the strategy of Crisostomi et al. (2024): for a single layer, we replace each permutation matrix with a doubly stochastic matrix Q, i.e., a *soft* assignment of neurons, and maximize the evidence with the Frank–Wolfe algorithm. To evaluate the evidence at a soft assignment, we apply Q to each client's diagonal precisions $\boldsymbol{\lambda}_k$ and precision-weighted means $\boldsymbol{\lambda}_k \odot \hat{\mathbf{w}}_k$. The merged precision and precision-weighted means are then linear in Q, so the evidence remains in closed form; it coincides with Equation (5) at permutation matrices. Each Frank–Wolfe iteration consists of two steps: (i) it finds the permutation that best agrees with the current gradient of the evidence, which is a linear assignment problem; (ii) it moves the soft assignment towards that permutation, with a step size chosen by line search. We sweep over the layers repeatedly, keeping the neighboring layers at their most recent assignments, and at convergence round each soft assignment to the closest permutation. Unlike Crisostomi et al. (2024), whose objective is defined over all pairs of models, we align every client to the same merged model, so the cost grows linearly, rather than quadratically, with the number of clients.

Global gauge. After optimization, applying the same permutation to all clients changes neither the evidence nor the function computed by the merged model, so the permutations are defined only up to such a common permutation (Appendix A.3). We therefore optimize the permutations of all clients jointly and, after convergence, compose them with the inverse permutation of the first client. This sets the first client's permutation to the identity without changing the merged function. Alternatively, one client can be fixed as a reference from the start, and only the permutations of the remaining $K-1$ clients optimized. The two choices give comparable results (see Section 5.3).

Bayesian merge and global model. The server applies the selected permutations to the clients' Laplace approximations and combines the aligned posteriors according to Equation (1). After the construction of the global posterior, the mode of this (Gaussian) distribution is returned to the clients as the global model.

Implementation considerations. The full Hessian requires storage quadratic in the number of parameters, which is impractical for modern neural networks. We therefore use a Kronecker-factored approximation of the Hessian for the clients' Laplace posteriors (Grosse & Martens, 2016; Ritter et al., 2018), and its diagonal for alignment. For the prior, we use a permutations-invariant zero-

mean isotropic Gaussian $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma_p^2 I)$ (Wilson & Izmailov, 2020), with $\sigma_p^2 = 0.1$ by default. Following Talpini et al. (2025), we use a tempered posterior, obtained by raising the likelihood to the power $1/T$ with $T = 0.1$. Each posterior is therefore $\Lambda_k = \frac{1}{T} H_k + \sigma_p^{-2} I$, where H_k is the Kronecker-factored curvature; the prior term is added exactly in the Kronecker eigenbasis, and the diagonal used for alignment is that of Λ_k. A sensitivity analysis on σ^2 and T is reported in Appendix D.1.

5 Experiments

5.1 Experimental Setting

We adopted the experimental setting of Jhunjhunwala et al. (2024); Talpini et al. (2025). We experimented on three datasets and two models: FashionMNIST Xiao et al. (2017), SVHN Netzer et al. (2011) and CIFAR-10 Krizhevsky et al. (2010), using LeNet LeCun et al. (1998) and a CNN proposed by Wang et al. (2020). Each training set is partitioned across clients by sampling label proportions from a Dirichlet distribution Hsu et al. (2019) with concentration $\alpha \in \{0.025, 0.05, 0.1, 0.2\}$, where smaller α means higher statistical heterogeneity; accuracy is always measured on the full centralized test set. Unless stated otherwise, clients are independently initialized and draw the number of their local epochs from a symmetric Beta-Binomial distribution with $n = 2T_e$ and $\alpha = \beta \equiv \eta$, $\eta = 1$, giving mean T_e and support $0, \ldots, 2T_e$. Any draw of 0 is replaced by 1. Here, $T_e = 20$ for LeNet and $T_e = 30$ for CNN, following Jhunjhunwala et al. (2024); further details are given in Appendix B. All main results are reported as mean $\pm$ standard error (SE) over 10 seeds.

5.1.1 Baselines

We compare FLAME with well-established state-of-the-art methods, selecting representative algorithms from each category described in Section 2. We focus on methods that neither require a dataset at the server nor involve multiple rounds of server–client communication. Specifically, we consider the model-merging methods **RegMean** (Jin et al., 2023) and **FisherMerge** (Matena & Raffel, 2022); **FedBEns** (Talpini et al., 2025), an ensemble approach implemented with two mixture components per client; the matching-based methods **PFNM** (Yurochkin et al., 2019) and **OTFusion** (Singh & Jaggi, 2020); and **FedAvg** (McMahan et al., 2017) as a standard reference.

We also include baselines that isolate the individual components of our pipeline. **FedFisher** (Jhunjhunwala et al., 2024), which we report as $\Pi{=}I$, performs the same Bayesian merge as FLAME without any alignment. $\boldsymbol{C^2M^3}$ Crisostomi et al. (2024) is the extension of Git Re-Basin in the case of multiple models: it aligns weights, enforcing Euclidean proximity and cycle-consistency. Moreover, we also consider **Cycle**, which first performs cycle-consistent alignment using the objective of Crisostomi et al. (2024), and then applies the same Bayesian merge as FLAME; in this way, we isolate the alignment criterion of FLAME. Implementation details for the baselines are given in Appendix B.2.

5.2 Main results

Uncoordinated clients. In this Section, we start by analyzing the performance of the considered methods in a fully uncoordinated setting, i.e., with different randomly initialized models' weights and different training epochs. We fix $K = 10$, while varying the heterogeneity parameter α. Table 1 reports the results: FLAME achieves the highest mean accuracy in all but one configuration. The largest margins are on SVHN: e.g., at $\alpha = 0.1$ it reaches 55.1%, against 37.5% for the unaligned Bayesian merge and 38.8% for Cycle. On CIFAR-10, gains are smaller, often within one SE.

Table 1 supports three observations. First, the choice of the alignment criterion matters: Cycle, which uses the same Bayesian merge as FLAME, is statistically indistinguishable from no alignment across most of the grid. Since Cycle differs from FLAME only in the alignment objective, Euclidean weight matching carries little signal for the merge once clients differ in their data. Second, alignment followed by curvature-agnostic aggregation is not enough: C^2M^3, OTFusion and PFNM performance remains far below FLAME. Third, curvature without alignment is not enough either: $\Pi{=}I$, FisherMerge and FedBEns weight parameters by their precision but implicitly assume a shared basin, and they also trail FLAME. Only the combination of evidence-based alignment and

Table 1: Test accuracy (%), 10 clients, random init, random epochs. Mean ± SE (10 seeds). Best in bold, second best underlined.

Mean Accuracy & Standard Error (10 seeds)										
Dataset	FLAME (Our)	$\Pi{=}I$	Cycle	C^2M^3	RegMean	FedAvg	Fisher Merge	FedBEns	PFNM	OTFusion
$\alpha = 0.025$										
FMNIST	**26.3±2.4**	22.2±2.3	24.9±1.4	10.6±0.6	12.3±1.1	9.8±0.4	19.1±2.5	22.4±2.1	12.5±1.6	11.0±1.0
SVHN	**34.2±2.7**	29.5±2.7	26.0±2.2	13.4±1.8	11.8±1.5	13.9±1.9	23.1±3.0	26.8±3.3	12.1±1.9	10.1±1.4
CIFAR-10	**22.9±2.1**	20.2±2.0	20.6±2.2	10.6±0.4	11.0±0.5	10.3±0.2	17.1±1.8	18.7±2.2	10.2±0.2	10.0±0.0
$\alpha = 0.05$										
FMNIST	**29.4±1.7**	26.7±1.8	23.3±1.5	12.0±1.0	19.2±2.5	10.0±0.0	15.9±1.5	19.6±1.8	13.4±1.2	16.6±2.6
SVHN	**46.8±2.3**	32.7±2.5	30.9±2.9	13.8±1.5	18.5±2.7	14.9±1.5	26.7±2.4	28.1±2.1	11.9±1.4	15.5±2.9
CIFAR-10	**27.6±1.1**	24.3±1.2	24.1±1.6	11.0±0.4	12.8±0.7	10.9±0.5	17.6±1.3	24.3±1.5	10.1±0.1	11.8±1.1
$\alpha = 0.1$										
FMNIST	**38.4±2.2**	29.1±3.0	27.8±3.1	11.3±0.9	20.3±1.9	10.4±0.2	15.1±0.9	24.6±3.3	12.5±1.2	31.4±4.0
SVHN	**55.1±3.6**	37.5±3.3	38.8±2.7	10.9±1.4	18.8±1.0	11.4±1.6	26.5±2.1	31.5±1.8	15.8±2.1	28.5±1.8
CIFAR-10	**26.9±1.7**	24.7±1.5	23.7±1.6	10.9±0.2	14.5±0.6	10.9±0.3	17.5±1.2	24.6±1.7	11.9±0.9	23.7±1.6
$\alpha = 0.2$										
FMNIST	42.9±2.6	33.0±2.7	30.7±1.5	11.0±0.7	24.6±2.8	10.7±0.5	14.2±1.6	26.7±1.8	18.2±1.9	**45.8±2.5**
SVHN	**65.5±1.6**	47.1±2.1	47.6±2.3	17.8±1.0	23.6±1.6	16.6±1.2	33.5±1.8	42.5±2.4	22.4±3.0	51.1±3.6
CIFAR-10	**35.7±1.0**	32.0±1.2	31.5±1.5	11.4±0.3	14.6±0.3	12.0±0.4	20.4±1.6	30.1±1.3	13.2±0.7	33.2±2.1

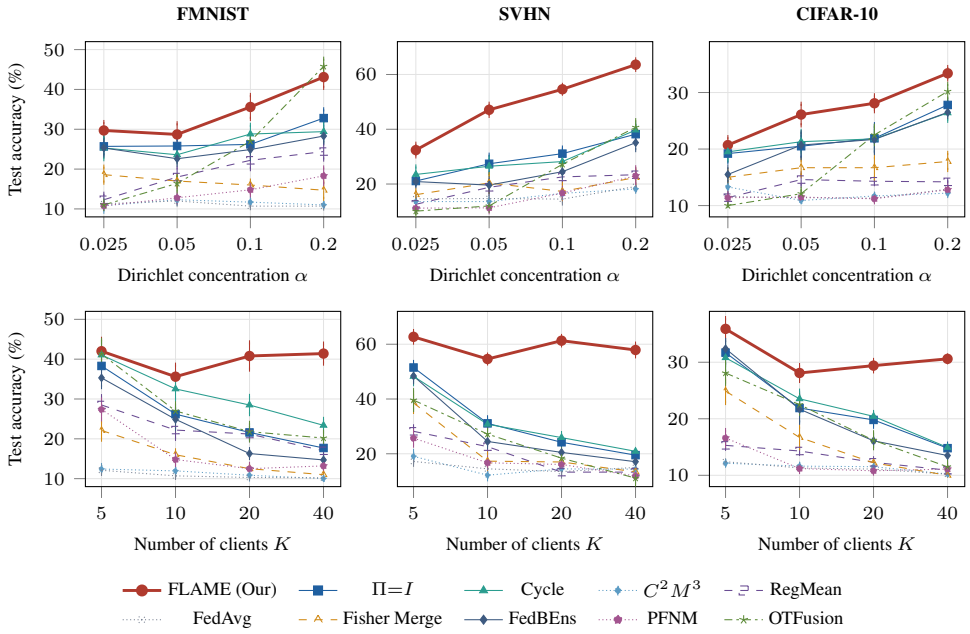

Figure 2: Test accuracy (%) with random initialization and the same number of epochs, by dataset (columns). Top row: varying the Dirichlet concentration α with 10 clients, where smaller α means more heterogeneous client data. Bottom row: varying the number of clients K at $\alpha = 0.1$. Markers report the mean over 10 seeds and bars report the standard error.

curvature-aware merging recovers a substantially better global model. The single exception is FashionMNIST at $\alpha = 0.2$, the mildest heterogeneity on the shallowest architecture, where OTFusion slightly outperforms FLAME. This is also the regime in which alignment has the least to correct.

In Appendix C, we analyze the effect of initialization and training epochs, and we report that the main real effect on model-space divergence comes mostly from differences in initialization. In the standard coordinated scenario, FLAME performs similarly to the unaligned Bayesian merge ($\Pi = I$) and Cycle (see Table 5, Appendix C.2), as we would expect from our assumptions.

Heterogeneity and number of participants. Figure 2 explores two dimensions that control the difficulty of the merge operation, namely client data heterogeneity and the number of clients, while keeping the other fixed. In this setting, clients are randomly initialized but trained for the same number of epochs, to better isolate the effects of data heterogeneity and the number of clients. Lower heterogeneity (top row) improves every method, but the ranking remains stable. The gap between FLAME and the other methods narrows at the most extreme heterogeneity ($\alpha = 0.025$), where all methods achieve low accuracy. When the number of participants varies at fixed $\alpha = 0.1$ (bottom row), the advantage of FLAME widens with K, while the baselines degrade as more clients are added. This is the behavior expected if misalignment is the dominant failure mode: each additional client contributes another set of arbitrarily permuted units, so the cost of merging in the wrong gauge compounds, whereas a correctly aligned merge keeps accumulating information.

5.3 Ablation studies

Table 2: Contribution of each term of the evidence objective. 10 clients, random init, random epochs, α=0.1. Mean $\pm$ SE (10 seeds).

	FMNIST	SVHN	CIFAR-10
None ($\Pi = I$)	29.1±3.0	37.5±3.3	24.7±1.5
Cycle	27.8±3.1	38.8±2.7	23.7±1.6
Misalignment	36.9±2.6	46.0±2.9	26.8±1.7
+ Volume/Occam	37.6±2.5	49.9±2.7	**27.0±1.7**
+ Prior-location (FLAME)	**38.4±2.2**	**55.1±3.6**	26.9±1.7

Table 3: Choice of the template. 10 clients, random init, random epochs, α=0.1. Mean $\pm$ SE (10 seeds).

	FMNIST	SVHN	CIFAR-10
Free (FLAME)	38.4±2.2	**55.1±3.6**	**26.9±1.7**
Highest ev.	36.1±3.0	54.0±3.5	26.7±1.9
Lowest ev.	37.9±2.7	54.4±3.6	26.2±1.5
Random	**38.9±2.9**	54.1±2.9	26.3±1.6
Π=I	29.1±3.0	37.5±3.3	24.7±1.5

Terms of the evidence objective. Table 2 adds the terms of Equation (5) one at a time, starting from no alignment and standard weight matching. The results show that Euclidean matching brings no gain over no alignment, whereas the misalignment term alone accounts for most of the improvement: measuring discrepancies with the curvature-induced metric is therefore the key ingredient. Moreover, the volume/Occam term provides a further, smaller gain, as does the prior-location term. Thus, each term either improves or preserves accuracy, and the full evidence performs overall best.

Global gauge choice. FLAME leaves all client permutations free during optimization and fixes the global gauge afterwards. Alternatively, one client can be elected as an anchor and the remaining $K - 1$ clients aligned to it. We consider three anchor-selection heuristics: highest evidence, lowest evidence, and random selection. Table 3 shows that all perform within one SE, with no consistent winner, and remain well above $\Pi = I$. Thus, FLAME's gains come from the evidence objective, rather than the gauge choice, and no client acts as a privileged anchor.

6 Conclusion

We framed alignment and aggregation in One-Shot FL as a single inference problem and proposed FLAME, which selects the permutations maximizing the Bayesian evidence of the merged model and merges under that alignment, so that a curvature-based quantity governs both steps. The evidence admits a closed form for any number of clients and contains the weight matching of Git Re-Basin as its isotropic special case. Empirically, the combination is what matters: Euclidean alignment adds little under data heterogeneity, while curvature-aware merging alone fails when the local posteriors lie in different regions of parameter space; only their proper combination recovers a better global model.

Limitations and future work. Neuron-wise decomposition of the evidence requires a diagonal posterior, and the benefit of alignment arises when clients start from different initializations. On more complex architectures and tasks, all methods, ours included, remain weak, suggesting that one-shot aggregation at depth requires more than alignment (see Appendix C.3). Natural extensions include heterogeneous architectures, other network symmetries, priors spanning all functionally equivalent parameterizations, and using the selected alignments as a lens on the role of individual parameters.

AI USE STATEMENT

In this work, we used generative AI tools for code generation, polishing and grammatical revision of the manuscript, and checking mathematical proofs. We have not used generative AI tools to generate experimental results, conduct experiments, make scientific claims, or make decisions about the research methodology, and the remaining required disclosure tasks are not applicable to this work.

All AI-assisted code was reviewed, adapted, and tested by the authors for correctness. AI-assisted text was reviewed and revised by the authors to ensure that it accurately reflects the intended content and scientific contributions. AI-assisted proof checking was independently verified by the authors, who checked the mathematical arguments and their consistency with the definitions and results in the paper. We take responsibility for the final content of this work, including all text, claims, mathematical arguments, code, and experimental results produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The full FLAME procedure is described in Section 4 and given as pseudo-code in Algorithm 1 (Appendix B.1), together with its computational cost. All theoretical results are stated with their assumptions in Section 3, and complete proofs are provided in Appendix A. The experimental protocol, including datasets, architectures, the Dirichlet partitioning of the data, the sampling of local epochs, and the number of random seeds, is described in Section 5; the implementation details of each baseline are given in Appendix B.2. A sensitivity analysis for the prior variance and the likelihood temperature is reported in Appendix D.1. Laplace approximations are computed with the open-source `laplace` library (Daxberger et al., 2021), and baselines experiments rely on the authors' official implementations where available. An anonymized version of our code, with scripts to reproduce all tables and figures, is included in the supplementary material, and it will be released as a public GitHub repository upon publication.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. In *International Conference on Learning Representations (ICLR)*, 2021.

Alberto Bernacchia. Global curvature for second-order optimization of neural networks. In *International Conference on Machine Learning (ICML)*, 2025.

Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodolà. c^2m^3: Cycle-consistent multi-model merging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 (NeurIPS)*, 2021.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Ensemble attention distillation for privacy-preserving federated learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning (ICML)*, 2016.

Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.

Mohsin Hasan, Guojun Zhang, Kaiyang Guo, Xi Chen, and Pascal Poupart. Calibrated one round federated learning with bayesian inference in the predictive space. In *AAAI Conference on Artificial Intelligence*, 2024.

Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. *Advanced Neural Computers*, pp. 129–135, 1990.

Clare Elizabeth Heinbaugh, Emilio Luz-Ricca, and Huajie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *International Conference on Learning Representations (ICLR)*, 2023.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. In *NeurIPS Workshop on Federated Learning*, 2019.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning (ICML)*, 2021.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedfisher: Leveraging fisher information for one-shot federated learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. Repair: Renormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations (ICLR)*, 2023.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). `http://www.cs.toronto.edu/~kriz/cifar.html`, 2010.

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.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chang Liu, Chenfei Lou, Runzhong Wang, Alan Yuhan Xi, Li Shen, and Junchi Yan. Deep neural network fusion via graph matching with applications to model ensemble and federated learning. In *International Conference on Machine Learning (ICML)*, 2022.

Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. Fedlpa: One-shot federated learning with layer-wise posterior aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Kevin P. Murphy. *Probabilistic machine learning: advanced topics*. MIT Press, 2023.

Anshul Nasery, Jonathan Hayase, Pang Wei Koh, and Sewoong Oh. Pleas—merging models with permutations and least squares. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat. Where to begin? on the impact of pre-training and initialization in federated learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Filippo Rinaldi, Aniello Panariello, Giacomo Salici, Angelo Porrello, and Simone Calderara. Transporting task vectors across different architectures without training. In *International Conference on Machine Learning (ICML)*, 2026.

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

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

George Stoica, Daniel Bolya, Jakob Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. ZipIt! merging models from different tasks without training. In *International Conference on Learning Representations (ICLR)*, 2024.

Jacopo Talpini, Marco Savi, and Giovanni Neglia. Fedbens: One-shot federated learning based on bayesian ensemble. In *International Conference on Machine Learning (ICML)*, 2025.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023.

Haibo Yang, Xin Zhang, Prashant Khanduri, and Jia Liu. Anarchic federated learning. In *International Conference on Machine Learning (ICML)*, 2022.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Hui Zeng, Wenke Huang, Tongqing Zhou, Xinyi Wu, Guancheng Wan, Yingwen Chen, and Zhiping Cai. Does one-shot give the best shot? mitigating model inconsistency in one-shot federated learning. In *International Conference on Machine Learning (ICML)*, 2025.

Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020.

A PROOFS

This Appendix collects the theory behind Section 4. Appendix A.1 fixes the notation and the two Gaussian identities used throughout. Appendix A.2 derives the evidence of the merged model for any number of clients (Theorem A.2), of which Proposition 4.1 is the two-client case. Appendix A.3 proves the gauge invariance that allows fixing one client's permutation. Appendices A.4 and A.5 study two special cases: the broad-prior limit, and isotropic posteriors, for which evidence maximization reduces to weight matching (Proposition 4.2).

A.1 SETTING, NOTATION AND GAUSSIAN IDENTITIES

We consider $K \geq 2$ clients and a network with d parameters. The prior is $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma_p^2 I)$, shared across clients. Client k summarizes its local posterior by the Laplace approximation $\mathcal{N}(\mathbf{a}_k, A_k)$, with precision

$$A_k^{-1} = \tfrac{1}{T} H_k + \sigma_p^{-2} I, \tag{8}$$

where $H_k \succeq 0$ is the curvature of the local negative log-likelihood (Kronecker-factored in practice) and T is the temperature. Tempering rescales only the likelihood, so Equation (1) holds with tempered likelihoods. By construction,

$$A_k^{-1} \succeq \sigma_p^{-2} I \qquad \text{for all } k. \tag{9}$$

Permuted posteriors. For permutations $\Pi_{1:K} = (\Pi_1, \ldots, \Pi_K) \in \mathcal{P}^K$, we write $\tilde{\mathbf{a}}_k = \Pi_k \mathbf{a}_k$ and $\tilde{A}_k = \Pi_k A_k \Pi_k^\top$ for the representative of client k in its orbit (Equation (4)). Their product is characterized by

$$C_\Pi = \Big(\sum_{k=1}^K \tilde{A}_k^{-1} \Big)^{-1}, \qquad \mathbf{c}_\Pi = C_\Pi \sum_{k=1}^K \tilde{A}_k^{-1} \tilde{\mathbf{a}}_k, \tag{10}$$

and the global posterior of Equation (1) is $\mathcal{N}(\boldsymbol{\mu}_\Pi, \Sigma_\Pi)$, with

$$\Sigma_\Pi^{-1} = C_\Pi^{-1} - (K-1)\, \sigma_p^{-2} I, \qquad \boldsymbol{\mu}_\Pi = \Sigma_\Pi\, C_\Pi^{-1} \mathbf{c}_\Pi. \tag{11}$$

By Equation (9), $C_\Pi^{-1} \succeq K \sigma_p^{-2} I$, hence $\Sigma_\Pi^{-1} \succeq \sigma_p^{-2} I \succ 0$ and the global posterior is well defined. The mode $\boldsymbol{\mu}_\Pi$ is the global model returned by FLAME. We omit the subscript Π when no ambiguity arises.

Lemma A.1 (Gaussian identities)**.**

(i) Product. $\prod_{k=1}^K \mathcal{N}(\mathbf{w}|\tilde{\mathbf{a}}_k, \tilde{A}_k) = \gamma_\Pi\, \mathcal{N}(\mathbf{w}|\mathbf{c}_\Pi, C_\Pi)$*, with*

$$\begin{aligned} \ln \gamma_\Pi &= -\tfrac{(K-1)d}{2} \ln(2\pi) + \tfrac{1}{2} \ln |C_\Pi| - \tfrac{1}{2} \sum_{k=1}^K \ln |\tilde{A}_k| - \tfrac{1}{2} R_\Pi, \\ R_\Pi &:= \sum_{k=1}^K (\tilde{\mathbf{a}}_k - \mathbf{c}_\Pi)^\top \tilde{A}_k^{-1} (\tilde{\mathbf{a}}_k - \mathbf{c}_\Pi). \end{aligned} \tag{12}$$

Moreover, $R_\Pi = \sum_k \tilde{\mathbf{a}}_k^\top \tilde{A}_k^{-1} \tilde{\mathbf{a}}_k - \mathbf{c}_\Pi^\top C_\Pi^{-1} \mathbf{c}_\Pi$*, and for two clients* $(\mathbf{a}, A)$*,* $(\Pi\mathbf{b}, \Pi B \Pi^\top)$*,* $R_\Pi = (\mathbf{a} - \Pi\mathbf{b})^\top (A + \Pi B \Pi^\top)^{-1} (\mathbf{a} - \Pi\mathbf{b})$.

(ii) Ratio. *For* $s^2 > 0$ *and* $C \prec s^2 I$*,*

$$\int \frac{\mathcal{N}(\mathbf{w}|\mathbf{c}, C)}{\mathcal{N}(\mathbf{w}|\mathbf{0}, s^2 I)}\, d\mathbf{w} = (2\pi)^{d/2} \frac{|s^2 I|}{|s^2 I - C|^{1/2}} \exp\Big(\tfrac{1}{2}\, \mathbf{c}^\top (s^2 I - C)^{-1} \mathbf{c} \Big). \tag{13}$$

Proof. (i) Completing the square, $\sum_k (\mathbf{w} - \tilde{\mathbf{a}}_k)^\top \tilde{A}_k^{-1} (\mathbf{w} - \tilde{\mathbf{a}}_k) = (\mathbf{w} - \mathbf{c})^\top C^{-1} (\mathbf{w} - \mathbf{c}) + \sum_k \tilde{\mathbf{a}}_k^\top \tilde{A}_k^{-1} \tilde{\mathbf{a}}_k - \mathbf{c}^\top C^{-1} \mathbf{c}$. Expanding R_Π and using $\sum_k \tilde{A}_k^{-1} \tilde{\mathbf{a}}_k = C^{-1} \mathbf{c}$ shows that the last two terms equal R_Π. Comparing the normalizing constants of the two sides gives Equation (12). The two-client form of R_Π is the standard product identity $\mathcal{N}(\mathbf{w}|\mathbf{a}, A)\, \mathcal{N}(\mathbf{w}|\mathbf{b}', B') = \mathcal{N}(\mathbf{a}|\mathbf{b}', A + B')\, \mathcal{N}(\mathbf{w}|\mathbf{c}, C)$ (Rasmussen & Williams, 2006).

(ii) Let $M := C^{-1} - s^{-2}I \succ 0$. The ratio equals $|s^2I|^{1/2}|C|^{-1/2}\exp\big(-\frac{1}{2}\mathbf{w}^\top M\mathbf{w} + \mathbf{w}^\top C^{-1}\mathbf{c} - \frac{1}{2}\mathbf{c}^\top C^{-1}\mathbf{c}\big)$, and the Gaussian integral gives $(2\pi)^{d/2}|s^2I|^{1/2}(|C||M|)^{-1/2}\exp\big(\frac{1}{2}\mathbf{c}^\top(C^{-1}M^{-1}C^{-1} - C^{-1})\mathbf{c}\big)$. Since $M = C^{-1}(I - s^{-2}C)$ and C commutes with $I - s^{-2}C$, we have $C^{-1}M^{-1}C^{-1} - C^{-1} = (s^2I - C)^{-1}$ and $|C||M| = |s^2I - C|/|s^2I|$, which gives Equation (13). □

A.2 Evidence of the merged model

Theorem A.2 (Evidence of the merged model)**.** *Let each client's evidence be given by the Laplace approximation of Equation (2),* $\ln p(\mathcal{D}_k) = \ln p(\mathcal{D}_k \mid \mathbf{a}_k) + \ln p(\mathbf{a}_k) + \frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|A_k|$*, and let each local posterior be represented by* $\mathcal{N}(\tilde{\mathbf{a}}_k, \tilde{A}_k)$ *when merged through Equation (1). Then the log-evidence of the merged model is*

$$\ln p(\mathcal{D}_{1:K} \mid \mathcal{M}_\Pi) = \sum_{k=1}^{K}\Big[\ln p(\mathcal{D}_k \mid \mathbf{a}_k) - \frac{\|\mathbf{a}_k\|^2}{2\sigma_p^2}\Big] + Z(\Pi_{1:K}), \tag{14}$$

where the first sum does not depend on the permutations and

$$Z(\Pi_{1:K}) = \underbrace{-\tfrac{1}{2}R_\Pi}_{\textit{Misalignment}} \underbrace{+\tfrac{1}{2}\ln|C_\Pi| - \tfrac{1}{2}\ln\big|\sigma_p^2 I - (K-1)C_\Pi\big|}_{\textit{Volume/Occam}} \underbrace{+\tfrac{K-1}{2}\,\mathbf{c}_\Pi^\top\big(\sigma_p^2 I - (K-1)C_\Pi\big)^{-1}\mathbf{c}_\Pi}_{\textit{Prior-location}} \tag{15}$$

Equation (14) holds exactly under the Laplace approximation: all constants cancel.

Proof. Normalization constant. Under conditional independence, $\prod_k p(\mathbf{w} \mid \mathcal{D}_k) = p(\mathcal{D}_{1:K} \mid \mathbf{w})\,p(\mathbf{w})^K / \prod_k p(\mathcal{D}_k)$, so the right-hand side of Equation (1) equals $\alpha\, p(\mathbf{w} \mid \mathcal{D}_{1:K})\,p(\mathcal{D}_{1:K}) / \prod_k p(\mathcal{D}_k)$. Hence $\alpha = \prod_k p(\mathcal{D}_k)/p(\mathcal{D}_{1:K})$ and

$$\ln p(\mathcal{D}_{1:K}) = \sum_{k=1}^{K}\ln p(\mathcal{D}_k) + \ln\alpha^{-1}, \qquad \alpha^{-1} = \int \frac{\prod_k p(\mathbf{w} \mid \mathcal{D}_k)}{p(\mathbf{w})^{K-1}}\,d\mathbf{w}. \tag{16}$$

Since each Π_k preserves the network function and the prior is isotropic, $\ln p(\mathcal{D}_k \mid \tilde{\mathbf{a}}_k) = \ln p(\mathcal{D}_k \mid \mathbf{a}_k)$, $\ln p(\tilde{\mathbf{a}}_k) = \ln p(\mathbf{a}_k)$ and $\ln|\tilde{A}_k| = \ln|A_k|$, so the client evidences do not depend on the choice of representative.

Computing α^{-1}. Write $p(\mathbf{w})^{K-1} = \kappa\,\mathcal{N}(\mathbf{w}|\mathbf{0}, s^2I)$ with $s^2 := \sigma_p^2/(K-1)$ and $\ln\kappa = \frac{d}{2}\ln(2\pi s^2) - \frac{(K-1)d}{2}\ln(2\pi\sigma_p^2)$. By Lemma A.1(i), $\alpha^{-1} = (\gamma_\Pi/\kappa)\int \mathcal{N}(\mathbf{w}|\mathbf{c}, C)/\mathcal{N}(\mathbf{w}|\mathbf{0}, s^2I)\,d\mathbf{w}$. By (9), $C \preceq \frac{\sigma_p^2}{K}I \prec s^2I$, so Lemma A.1(ii) applies. Using $s^2I - C = \big(\sigma_p^2 I - (K-1)C\big)/(K-1)$:

$$\begin{aligned}\ln\alpha^{-1} = {} & \ln\gamma_\Pi - \ln\kappa + \tfrac{d}{2}\ln(2\pi) + d\ln s^2 + \tfrac{d}{2}\ln(K-1) + \\ & - \tfrac{1}{2}\ln\big|\sigma_p^2 I - (K-1)C\big| + \tfrac{K-1}{2}\,\mathbf{c}^\top\big(\sigma_p^2 I - (K-1)C\big)^{-1}\mathbf{c}.\end{aligned}$$

Collecting terms. Substitute this expression, the client evidences, and $\ln p(\mathbf{a}_k) = -\frac{d}{2}\ln(2\pi\sigma_p^2) - \|\mathbf{a}_k\|^2/(2\sigma_p^2)$ into Equation (16), and expand $\ln\gamma_\Pi$ by Equation (12). The terms $\frac{1}{2}\ln|A_k|$ from the client evidences cancels with $-\frac{1}{2}\ln|\tilde{A}_k|$ from $\ln\gamma_\Pi$. In addition, the coefficients of $\ln(2\pi)$, $\ln\sigma_p^2$ and $\ln(K-1)$ also sum to zero (using $\ln s^2 = \ln\sigma_p^2 - \ln(K-1)$). The remaining terms are Equation (14). □

Corollary A.3 (Two clients)**.** *For* $K = 2$*, with* $\Pi_1 = I$*,* $\Pi_2 = \Pi$ *and clients* $(\mathbf{a}, A)$*,* $(\mathbf{b}, B)$*, Equation (15) equals Equation (5) up to a constant independent of* Π*. This proves Proposition 4.1.*

Proof. The misalignment term follows from the two-client form of R_Π in Lemma A.1(i). For the volume term, $C_\Pi^{-1} = A^{-1}(A + \Pi B\Pi^\top)\,\Pi B^{-1}\Pi^\top$ gives $\frac{1}{2}\ln|C_\Pi| = -\frac{1}{2}\ln|A + \Pi B\Pi^\top| + \frac{1}{2}\ln|A| + \frac{1}{2}\ln|B|$, where the last two terms do not depend on Π. □

The misalignment term of Equation (15) is the sum of each client's Mahalanobis distance to the product mean $\mathbf{c}_\Pi$, whereFLAME aligns every client to (Section 4.2). The two log-determinants have a direct interpretation.

Lemma A.4 (Volume term as the global-posterior volume)**.** $\frac{1}{2}\ln|C_\Pi| - \frac{1}{2}\ln\left|\sigma_p^2 I - (K-1)C_\Pi\right| = \frac{1}{2}\ln|\Sigma_\Pi| - \frac{d}{2}\ln\sigma_p^2$.

Proof. By Equation (11), $\Sigma_\Pi^{-1} = C_\Pi^{-1}\big(I - (K-1)\sigma_p^{-2}C_\Pi\big)$, hence $\ln|\Sigma_\Pi| = \ln|C_\Pi| - \ln\left|\sigma_p^2 I - (K-1)C_\Pi\right| + d\ln\sigma_p^2$. □

Remark (compact form). Similarly, one can check that the prior-location term equals $\frac{1}{2}\big(\boldsymbol{\mu}_\Pi^\top\Sigma_\Pi^{-1}\boldsymbol{\mu}_\Pi - \mathbf{c}_\Pi^\top C_\Pi^{-1}\mathbf{c}_\Pi\big)$. Together with Lemmas A.1(i) and A.4, this gives $Z(\Pi_{1:K}) = \frac{1}{2}\boldsymbol{\eta}_\Pi^\top\Sigma_\Pi\boldsymbol{\eta}_\Pi + \frac{1}{2}\ln|\Sigma_\Pi| + \text{const}$, with $\boldsymbol{\eta}_\Pi = \sum_k \tilde{A}_k^{-1}\tilde{\mathbf{a}}_k$: the evidence depends on the permutations only through the global posterior. The three-term decomposition of Equation (15) is therefore one of several equivalent ways to write Z; we use it for interpretation and for the ablation of Table 2.

A.3 Permutation invariance and gauge fixing

Proposition A.5 (Gauge invariance)**.** *For any $\Pi_0 \in \mathcal{P}$, applying Π_0 to all clients gives*

$$\mathbf{c}_{\Pi_0\Pi} = \Pi_0\,\mathbf{c}_\Pi, \quad C_{\Pi_0\Pi} = \Pi_0\,C_\Pi\,\Pi_0^\top, \quad \boldsymbol{\mu}_{\Pi_0\Pi} = \Pi_0\,\boldsymbol{\mu}_\Pi, \quad \Sigma_{\Pi_0\Pi} = \Pi_0\,\Sigma_\Pi\,\Pi_0^\top, \tag{17}$$

and $Z(\Pi_0\Pi_1, \ldots, \Pi_0\Pi_K) = Z(\Pi_1, \ldots, \Pi_K)$. The merged models $\boldsymbol{\mu}_{\Pi_0\Pi}$ and $\boldsymbol{\mu}_\Pi$ therefore represent the same function, and one may fix $\Pi_1 = I$ without loss of generality. For $K = 2$, the merged model depends only on the relative permutation $\Pi_1^\top\Pi_2$.

Proof. Since Π_0 is orthogonal, $(\Pi_0\tilde{A}_k\Pi_0^\top)^{-1} = \Pi_0\tilde{A}_k^{-1}\Pi_0^\top$, so $\sum_k \tilde{A}_k^{-1} \mapsto \Pi_0\big(\sum_k \tilde{A}_k^{-1}\big)\Pi_0^\top$ and $\sum_k \tilde{A}_k^{-1}\tilde{\mathbf{a}}_k \mapsto \Pi_0\sum_k \tilde{A}_k^{-1}\tilde{\mathbf{a}}_k$. This gives the transformation of $(\mathbf{c}, C)$ and, since $\Pi_0 I \Pi_0^\top = I$, of $(\boldsymbol{\mu}, \Sigma)$ through Equation (11). Each term of Equation (15) is a quadratic form or a log-determinant of matrices that transform by the same orthogonal change of basis, and is therefore unchanged. Since $\Pi_0 \in \mathcal{P}$ preserves the network function, $\boldsymbol{\mu}_{\Pi_0\Pi}$ and $\boldsymbol{\mu}_\Pi$ represent the same function. Choosing $\Pi_0 = \Pi_1^\top$ sets the first client's permutation to the identity, and for $K = 2$ leaves $\Pi_1^\top\Pi_2$ as the only free variable. □

A.4 Broad-prior limit

Let $\Delta(\Pi) := -\frac{1}{2}\ln\left|\sigma_p^2 I - (K-1)C_\Pi\right| + \frac{K-1}{2}\,\mathbf{c}_\Pi^\top\big(\sigma_p^2 I - (K-1)C_\Pi\big)^{-1}\mathbf{c}_\Pi$ collect the prior-dependent terms of Equation (15). The broad-prior limit is taken with the data fixed, i.e., with H_k and $\mathbf{a}_k$ held constant in Equation (8).

Lemma A.6 (Broad-prior reduction)**.** *Assume $H_k \succ 0$ for all k (e.g., by damping). As $\sigma_p^2 \to \infty$ with H_k and $\mathbf{a}_k$ fixed, $\Delta(\Pi) = -\frac{d}{2}\ln\sigma_p^2 + O(\sigma_p^{-2})$ for every $\Pi \in \mathcal{P}^K$. Hence $\Delta(\Pi) - \Delta(\Pi') \to 0$ for all $\Pi, \Pi' \in \mathcal{P}^K$.*

Proof. By Equation (8), $A_k \to TH_k^{-1}$, so C_Π and $\mathbf{c}_\Pi$ converge to finite limits and are bounded. Then $\big(\sigma_p^2 I - (K-1)C_\Pi\big)^{-1} = \sigma_p^{-2}I + O(\sigma_p^{-4})$, so the prior-location term is $O(\sigma_p^{-2})$, and $\ln\left|\sigma_p^2 I - (K-1)C_\Pi\right| = d\ln\sigma_p^2 + \ln\left|I - (K-1)C_\Pi/\sigma_p^2\right| = d\ln\sigma_p^2 + O(\sigma_p^{-2})$. □

Corollary A.7 (Broad-prior form of the objective)**.** *Under the assumptions of Lemma A.6, let $Z^\infty(\Pi) := -\frac{1}{2}R_\Pi^\infty + \frac{1}{2}\ln|C_\Pi^\infty|$ be computed with the data-only covariances $A_k^\infty = TH_k^{-1}$. Then $Z(\Pi) - Z(\Pi') \to Z^\infty(\Pi) - Z^\infty(\Pi')$ for all Π, Π' and, since $\mathcal{P}^K$ is finite, for σ_p^2 large enough every maximizer of Z maximizes Z^∞. For two clients,*

$$Z^\infty(\Pi) = -\tfrac{1}{2}\,\|\mathbf{a} - \Pi\mathbf{b}\|^2_{M(\Pi)} - \tfrac{1}{2}\ln\left|A^\infty + \Pi B^\infty\Pi^\top\right| + \text{const}, \qquad M(\Pi) := \big(A^\infty + \Pi B^\infty\Pi^\top\big)^{-1}. \tag{18}$$

Proof. By Lemma A.6, differences of Δ vanish, and the remaining terms of Z are continuous in A_k and converge to those of Z^∞. If Π does not maximize Z^∞, its gap $\delta > 0$ to the maximum is bounded away from zero because $\mathcal{P}^K$ is finite; for σ_p^2 large enough, all differences of Z are within $\delta/2$ of those of Z^∞, so Π cannot maximize Z. The two-client form follows from Corollary A.3. □

In this limit, alignment minimizes a squared Mahalanobis distance in which the curvature-induced matrix $M(\Pi)$ replaces the Euclidean inner product, traded off against the volume of the merged posterior.

A.5 Isotropic posteriors: weight matching

Proposition A.8 (Weight matching as the isotropic case)**.** *Let the local posteriors be isotropic, $A_k = \sigma_k^2 I$ (equivalently, $H_k \propto I$), and let $\lambda := \sum_k \sigma_k^{-2}$. Then, up to terms independent of the permutations,*

$$Z(\Pi_{1:K}) = \sum_{k<l} \beta_{kl} \langle \Pi_k \mathbf{a}_k, \Pi_l \mathbf{a}_l \rangle, \qquad \beta_{kl} = \frac{\gamma}{\sigma_k^2 \sigma_l^2}, \qquad \gamma = \frac{1}{\lambda} + \frac{K-1}{\lambda\left(\lambda\sigma_p^2 - (K-1)\right)} > 0. \tag{19}$$

Proof. Since Π_k is orthogonal, $\tilde{A}_k = \sigma_k^2 I$, so $C_\Pi = \lambda^{-1} I$ and both log-determinants in Equation (15) are constant. Moreover, $\mathbf{c}_\Pi = \lambda^{-1}\mathbf{u}$ with $\mathbf{u} := \sum_k \tilde{\mathbf{a}}_k/\sigma_k^2$. By Lemma A.1(i), $R_\Pi = \sum_k \|\mathbf{a}_k\|^2/\sigma_k^2 - \|\mathbf{u}\|^2/\lambda$, and the prior-location term equals $\frac{K-1}{2}\|\mathbf{u}\|^2/\big(\lambda(\lambda\sigma_p^2-(K-1))\big)$. Hence $Z = \frac{\gamma}{2}\|\mathbf{u}\|^2 + \text{const}$, and expanding $\|\mathbf{u}\|^2 = \sum_k \|\mathbf{a}_k\|^2/\sigma_k^4 + 2\sum_{k<l}\langle \tilde{\mathbf{a}}_k, \tilde{\mathbf{a}}_l\rangle/(\sigma_k^2\sigma_l^2)$ gives Equation (19). By Equation (9), $\lambda \geq K/\sigma_p^2$, so $\lambda\sigma_p^2 - (K-1) \geq 1$ and $\gamma > 0$. □

Corollary A.9 (Two clients; proof of Proposition 4.2)**.** *For $K = 2$, $A = \sigma_A^2 I$ and $B = \sigma_B^2 I$, $Z(\Pi) = \beta \langle \mathbf{a}, \Pi\mathbf{b}\rangle + \text{const}$ with*

$$\beta = \frac{1}{\sigma_A^2 + \sigma_B^2} + \frac{\omega_A\,\omega_B}{\sigma_p^2 - s^2} > 0, \qquad s^2 = \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2 + \sigma_B^2}, \quad \omega_A = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2}, \quad \omega_B = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2}. \tag{20}$$

Since $\|\Pi\mathbf{b}\|_2 = \|\mathbf{b}\|_2$, $\arg\max_\Pi Z(\Pi) = \arg\max_\Pi \langle \mathbf{a}, \Pi\mathbf{b}\rangle = \arg\min_\Pi \|\mathbf{a} - \Pi\mathbf{b}\|_2^2$.

Proof. Set $K = 2$ in Proposition A.8: $\lambda = 1/s^2$ and $\beta = \gamma/(\sigma_A^2\sigma_B^2)$, which simplifies to the stated expression. The last claim follows from $\|\mathbf{a} - \Pi\mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - 2\langle \mathbf{a}, \Pi\mathbf{b}\rangle$. □

For equal scales $\sigma_A = \sigma_B = \sigma$, $\beta = \frac{1}{2\sigma^2} + \frac{1}{4(\sigma_p^2 - \sigma^2/2)}$. With equal scales and $K > 2$, all β_{kl} coincide and Equation (19) is the unweighted sum of pairwise weight-matching objectives used for multi-model alignment.

B Additional experimental details

In this Appendix we provide the detailed FLAME pseudo-code with a computational analysis (Appendix B.1), and some additional information about the considered baselines (Appendix B.2).

B.1 FLAME pseudo-code

Algorithm 1 gives the full procedure described in Section 4.2. Lines 1–4 are executed by the clients and require a single transmission; everything that follows runs at the server. `PermAlign` implements the block-coordinate Frank–Wolfe search over the permutation groups, and `ServerMAP` the Bayesian merge under the selected alignment. To find the overall global model (i.e., the maximum of the posterior given in Equation (1)), the server performs $T_s = 100$ steps of Adam on the negative log-posterior. Note that this final merge uses the permuted Kronecker-factored Hessians, in line with Jhunjhunwala et al. (2024); Talpini et al. (2025), rather than the diagonal approximation used for alignment. To perform Laplace approximation of the local posteriors, we exploited the open-source `laplace` package Daxberger et al. (2021), a PyTorch library offering access to the most common Laplace approximations functionalities.

Algorithm 1 FLAME

Require: $\mathcal{D}_{1:K}$; server lr η_s and T_s steps; FW iters T_{FW}.
Uncoordinated setting: each client initialize the model independently with a random $\mathbf{w}_k^0$ and train it for different T_k steps
Coordinated setting: all clients share the same initial weights: $\mathbf{w}_k^0 = \mathbf{w}^0$ for all K clients, as well as same training epochs $T_k = T_e$

for client $k \in [K]$ **in parallel do**
  $\hat{\mathbf{w}}^{(k)}, \mathbf{\Lambda}^{(k)} \leftarrow$ TRAIN & LAPLACE$(\mathbf{w}_k^0, \mathcal{D}_k, T_k)$
  Send $\hat{\mathbf{w}}^{(k)}, \mathbf{\Lambda}^{(k)}$ to server
end for
$\mathbf{\Pi}_{1:K} \leftarrow$ PERMALIGN$(\hat{\mathbf{w}}^{1:K}, \mathrm{diag}(\mathbf{\Lambda}^{1:K}), T_{FW})$
Regauge: $\Pi_p^{(k)} \leftarrow (\Pi_p^{(1)})^{-1} \circ \Pi_p^{(k)} \; \forall\, k, p$ ▷ client 1 → identity
Apply $\Pi^{(k)}$ to $\hat{\mathbf{w}}^{(k)}, \mathbf{\Lambda}^{(k)} \to \tilde{\mathbf{w}}^{(k)}, \tilde{\mathbf{\Lambda}}^{(k)}$ for $k \geq 2$
$\hat{\mathbf{w}}^{\mathrm{global}} \leftarrow$ SERVERMAP$(\tilde{\mathbf{w}}^{1:K}, \tilde{\mathbf{\Lambda}}^{1:K}, T_s)$

function PERMALIGN$(\hat{\mathbf{w}}^{1:K}, \mathbf{\Lambda}^{1:K}, T_{FW})$
  Init $\mathbf{Q}_p^{(k)} \leftarrow \mathbf{I} \; \forall\, k \in [K]$, perm. group p
  for $t = 0, \ldots, T_{FW} - 1$ **do**
    for each group p **(Gauss-Seidel) do**
      Compute FW gradient $\mathbf{G}_p^{(k)}$ of $\log Z$ w.r.t. $\mathbf{Q}_p^{(k)}$ for all k (other groups fixed)
      $\mathbf{S}_p^{(k)} \leftarrow \arg\max_{\mathbf{S} \in \mathcal{P}} \langle \mathbf{S}, \mathbf{G}_p^{(k)} \rangle$ ▷ lin. assign.
      $\alpha^* \leftarrow \arg\max_{\alpha \in [0,1]} \ln Z_p(\mathbf{Q} + \alpha(\mathbf{S} - \mathbf{Q}))$
      $\mathbf{Q}_p^{(k)} \leftarrow \mathbf{Q}_p^{(k)} + \alpha^* (\mathbf{S}_p^{(k)} - \mathbf{Q}_p^{(k)})$
    end for
  end for
  Round $\mathbf{Q}_p^{(k)} \to \Pi_p^{(k)}$ via linear assignment
  return $\mathbf{\Pi}_{1:K}$
end function

function SERVERMAP$(\tilde{\mathbf{w}}^{1:K}, \tilde{\mathbf{\Lambda}}^{1:K}, T_s)$
  Build server objective $\mathcal{L}(\mathbf{w} \mid \tilde{\mathbf{w}}^{1:K}, \tilde{\mathbf{\Lambda}}^{1:K})$ taking the negative log of Eq. (1)
  $\mathbf{w}_s^0 \leftarrow$ weighted-mean$(\tilde{\mathbf{w}}^{1:K}, \mathrm{diag}(\tilde{\mathbf{\Lambda}}^{1:K}))$
  for $t = 0, \ldots, T_s - 1$ **do**
    $\mathbf{w}_s^{t+1} \leftarrow \mathrm{Adam}(\mathbf{w}_s^t, \nabla \mathcal{L}(\mathbf{w}_s^t), \eta_s)$
  end for
  return $\mathbf{w}_s^{T_s}$
end function

Scalability and computational cost. Communication and client computation are unchanged with respect to curvature-aware one-shot baselines like Jhunjhunwala et al. (2024): each client trains once and transmits its MAP estimate together with the Kronecker factors of $\Lambda^{(k)}$ in a single round. The additional cost of FLAME is confined to the server. Under the diagonal approximation, the evidence and its gradient are sums of per-parameter contributions and cost $O(Kd)$ with d the number of parameters; each Frank–Wolfe iteration adds one linear assignment problem per client and permutation group, of cost $O(d_\ell^3)$ for a layer of width d_ℓ, which is negligible for convolutional layers where d_ℓ is the number of channels. With P groups and T_{FW} sweeps, alignment therefore costs $O\big(T_{FW} P (Kd + K \max_\ell d_\ell^3)\big)$, linear in the number of clients: because every client is aligned to the merged posterior rather than to every other client, we avoid the quadratic scaling in K of pairwise objectives Crisostomi et al. (2024).

B.2 BASELINES

We compared our method with well-established state-of-the-art approaches for One-Shot Federated Learning, ensuring that we included representative algorithms from each category described in Section 2.1, with a special focus on those that do not explicitly require a dataset to be available at the server or involve multiple server-client communications.

Overall, we used the same basic training procedure for all the approaches and experiments to compare them fairly. SGD is exploited as local optimizer, with each client training for 20 epochs on LeNet and 30 epochs on the more complex CNNs for SVHN and CIFAR-10, in the same clients epochs setting. Moreover, we set the batch size to 64, the learning rate to 0.01, and the momentum to 0.9, in line with Jhunjhunwala et al. (2024); Talpini et al. (2025). For the various parameters of each baseline, we used their default settings and the official implementations provided by the authors. Regarding the experiments on the random initializations, we used the standard PyTorch weights initializer. In the following we briefly describe the considered baselines:

RegMean Jin et al. (2023) is a model fusion approach that merges models in their parameter space, guided by weights that minimize prediction differences between the merged model and the individual models.
Fisher Merge Matena & Raffel (2022) is a model fusion baseline that, similarly to Jhunjhunwala et al. (2024), employs a diagonal approximation to the Hessian of the log-posterior.
FedBEns Talpini et al. (2025) models the local posteriors using a mixture of Laplace approximations and aggregates these distributions at the server to account for the multimodality of the local loss landscapes. It then employs an ensemble for prediction as an approximate marginalization technique. In our experiments, we considered Kronecker-factored Hessians and two mixture components. Therefore, on the client side, the computational cost is approximately twice higher than that of our approach, as is the communication cost.
PNFM Yurochkin et al. (2019) aligns neurons across independently trained models by representing neurons through prototypes, and matching them using a Bayesian nonparametric approach. This allows functionally similar neurons to be identified despite differences in their ordering.
OTFusion Singh & Jaggi (2020) is a popular neuron-matching method that utilizes optimal transport to align neurons across the models by averaging their associated parameters, rather than directly averaging the weights.
FedFisher Jhunjhunwala et al. (2024) is based on an unimodal approximation of the loss as a quadratic function. The Fisher matrix is approximated through a diagonal matrix or through a Kronecker factorization. Here, we consider for comparison the method based on Kronecker approximation, referred to as FedFisher K-FAC, which achieves better results.
$\boldsymbol{C^2M^3}$Crisostomi et al. (2024) is a data-free model-merging method that jointly optimizes neuron permutations across layers and enforces cycle consistency when merging multiple models. The resulting aligned models are then merged in parameter space.

C Additional results

In this Appendix, we present additional results across different settings. We first examine the effect of random initialization and heterogeneous local training durations on global model performance (Appendix C.1). We then focus on the vanilla setting, with the same initialization and number of local epochs across clients (Appendix C.2). Finally, we evaluate a substantially more challenging setting using ResNet-18 on CIFAR-100, with fully uncoordinated initialization and training duration (Appendix C.3).

C.1 On the effect of random initialization and epochs

Table 4: Initialization × epochs scheme ($\alpha = 0.1$, $K = 10$, $\eta = 1$). Mean ± SE (5 seeds).

		Fixed epochs			**Random epochs** ($\eta = 1$)		
	Method	FMNIST	SVHN	CIFAR-10	FMNIST	SVHN	CIFAR-10
Rand init	**FLAME (Our)**	35.6 ± 3.0	54.6 ± 1.6	28.1 ± 1.4	38.6 ± 2.4	54.2 ± 3.8	27.3 ± 2.1
	Cycle	28.8 ± 2.3	28.2 ± 2.6	21.8 ± 1.8	26.9 ± 3.4	39.3 ± 1.9	22.7 ± 1.9
	$\Pi = I$	26.2 ± 1.9	31.1 ± 2.2	21.9 ± 2.5	29.3 ± 3.6	33.3 ± 4.6	23.1 ± 1.3
	C^2M^3	11.7 ± 0.9	16.6 ± 2.0	11.7 ± 0.4	10.9 ± 1.1	10.8 ± 1.5	10.9 ± 0.3
	OTFusion	27.0 ± 1.9	27.2 ± 2.8	22.5 ± 1.6	29.6 ± 5.7	29.3 ± 2.3	21.7 ± 2.2
Shared init	**FLAME (Our)**	62.6 ± 0.7	70.9 ± 1.2	34.7 ± 4.0	63.8 ± 2.1	60.3 ± 2.5	33.9 ± 1.0
	Cycle	62.7 ± 1.3	70.1 ± 1.6	35.7 ± 3.8	63.3 ± 2.1	61.2 ± 2.5	33.9 ± 1.0
	$\Pi = I$	62.4 ± 1.0	70.0 ± 1.6	34.3 ± 4.7	63.3 ± 2.1	61.7 ± 2.5	34.0 ± 1.1
	C^2M^3	43.8 ± 2.7	41.5 ± 7.7	26.5 ± 1.5	43.0 ± 3.6	39.9 ± 6.5	27.3 ± 2.2
	OTFusion	46.8 ± 1.4	45.5 ± 6.0	27.1 ± 1.8	44.3 ± 4.1	42.4 ± 5.0	28.8 ± 1.6

Table 4 crosses shared vs. random initialization with fixed vs. random epochs ($\eta = 1$), for $K = 10$ and $\alpha = 0.1$, over five seeds. Initialization is the dominant factor: with fixed epochs, a random initialization lowers $\Pi{=}I$ by roughly 36 p.p. on FashionMNIST, 39 on SVHN and by 12 on CIFAR-10. FLAME recovers part of this gap (23.5 p.p. on SVHN, 6.2 on CIFAR-10, 9.4 on FashionMNIST), while Cycle stays close to $\Pi{=}I$. With a shared initialization, FLAME, $\Pi{=}I$ and Cycle agree within one standard error, and parameter averaging remains about up to 16-29 p.p. behind. Random epochs have a smaller effect, which is mainly visible on SVHN.

Table 5: Test accuracy (%), 10 clients, shared init, same epochs. Mean ± SE (5 seeds).

Mean Accuracy & Standard Error (5 seeds)										
Dataset	FLAME (Our)	Π=I	Cycle	C^2M^3	RegMean	FedAvg	Fisher Merge	FedBEns	PFNM	OTFusion
					$\alpha = 0.025$					
FMNIST	47.4±4.1	**48.2±4.0**	47.4±3.9	19.8±3.7	18.2±4.8	19.9±3.8	32.6±6.3	36.8±5.3	18.8±8.8	27.6±4.8
SVHN	**47.4±3.5**	47.0±3.8	47.0±3.9	18.8±3.5	18.5±6.5	18.7±3.4	30.8±4.4	44.2±5.7	12.0±2.2	18.9±3.2
CIFAR-10	**25.9±2.2**	25.0±2.2	24.7±2.3	15.0±1.9	14.4±2.5	14.9±1.9	20.8±2.0	23.6±2.5	10.1±0.1	14.8±2.1
					$\alpha = 0.05$					
FMNIST	49.9±3.0	50.4±2.8	50.7±2.7	29.3±3.5	30.6±5.4	29.6±3.5	39.1±2.5	**55.5±3.0**	24.0±4.2	33.8±7.0
SVHN	60.4±1.9	**61.0±2.0**	59.2±2.0	28.7±3.5	51.8±5.4	28.6±3.5	39.0±5.5	59.1±3.6	15.3±2.3	28.2±2.0
CIFAR-10	**31.7±2.2**	30.7±2.3	30.9±2.5	22.8±1.2	31.6±1.2	23.0±1.1	25.5±2.0	29.9±3.3	11.4±1.2	20.9±1.5
					$\alpha = 0.1$					
FMNIST	62.6±0.7	62.4±1.0	62.7±1.3	43.8±2.7	50.3±2.8	43.7±2.7	49.5±5.8	**63.7±3.0**	41.5±2.2	46.8±1.4
SVHN	**70.9±1.2**	70.0±1.6	70.1±1.6	41.5 ± 7.7	58.8±4.0	41.5±7.7	52.4±3.0	70.0±2.4	14.4±3.0	45.5±6.0
CIFAR-10	34.7±4.0	34.3±4.7	35.7±3.8	26.5±1.5	33.9±1.1	26.5±1.4	26.8±3.9	**36.8±2.3**	10.6±0.5	27.1±1.8
					$\alpha = 0.2$					
FMNIST	73.6±2.2	73.0±2.6	73.0±2.4	52.8±4.2	62.4±2.2	52.9±4.2	68.3±2.1	**74.6±2.4**	52.8±3.1	55.0±4.5
SVHN	76.9±1.0	76.8±1.0	76.1±1.2	58.9±2.8	72.6±1.3	59.3±2.4	66.6±2.4	**80.4±0.8**	20.7±4.8	59.6±2.3
CIFAR-10	44.8±1.4	45.0±1.5	43.9±2.0	34.4±1.2	41.1±0.6	34.6±1.2	37.1±3.2	**48.1±0.6**	16.2±1.5	35.4±1.2

C.2 The standard FL setting: same init and same epochs

Table 5 repeats the grid of Table 1 with coordinated clients. FLAME, Π=I and Cycle agree within one standard error everywhere, while C^2M^3 and OTFusion align to FedAvg, indicating that matching tends to recover the identity: alignment has nothing to correct, and FLAME does not degrade performance in this setting. Curvature-aware merges outperform parameter averaging by 8 to 30 p.p. FedBEns has the highest mean in 6 out of 12 configurations, by at most 5.6 p.p. over FLAME, at twice the computation and communication costs (Appendix B.2). Accuracy remains well above the uncoordinated setting (e.g., 70.9% vs. 55.1% best on SVHN at $\alpha = 0.1$).

C.3 A harder setting: CIFAR-100 and ResNet-18

Table 6 repeats a similar setting of Section 5.2 on a deeper residual architecture. All methods, ours included, remain close to the random-guess level, and no alignment strategy recovers a usable model: the difficulty at this depth is not one that permutation alignment alone resolves. We report the grid for completeness and as a delimitation of the regime in which the method is useful. In addition, we emphasize that the considered setting is challenging for one-shot merging in general: each client observes only a small subset of the 100 classes, and even under a shared initialization, merging reaches only 10% accuracy. All curvature-aware methods perform comparably, indicating that at this scale the bottleneck lies in one-shot aggregation itself rather than in alignment.

Table 6: Test accuracy (%) on CIFAR-100 with ResNet-18, 5 clients, random initialization, random local epochs from Beta-Binomial with $T_e = 30$. Mean ± SE over 5 seeds.

α	FLAME (Our)	Π=I	Cycle	C^2M^3	FedAvg	Fisher Merge	FedBEns	OTFusion
0.025	7.4±0.5	**7.6±0.6**	6.9±0.7	1.0±0.0	1.0±0.0	1.1±0.1	7.1±0.5	2.3±0.2
0.05	7.5±0.6	**7.6±0.9**	7.4±1.0	1.0±0.1	1.0±0.0	1.2±0.1	7.4±1.0	2.4±0.3
0.1	**8.7±0.9**	7.8±1.4	8.2±1.1	1.0±0.0	1.0±0.0	1.4±0.2	7.0±1.4	1.9±0.2
0.2	**9.9±0.4**	8.0±1.4	8.1±1.3	1.0±0.0	1.0±0.0	1.1±0.0	7.7±1.1	2.9±0.4

D Additional ablations

In this Appendix we include additional ablation studies related to the likelihood temperature and prior variance (Appendix D.1), and to heterogeneity on the number of clients' epochs (Appendix D.2).

D.1 Temperature and prior sensitivity

The Laplace posteriors depend on two hyperparameters: the likelihood temperature T and the prior variance σ_p^2. Table 7 sweeps $T \in 0.01, 0.1, 1$ and $\sigma_p^2 \in 0.01, 0.1, 1, 10$ with ten clients, random initialization, random local epochs, and $\alpha = 0.1$, over five seeds. FLAME generally improves over $\Pi = I$ across the tested hyperparameter values, with the largest gains observed on SVHN. The relative improvement is less pronounced on FashionMNIST and CIFAR-10, where several configurations yield comparable performance between the two methods. For $T = 1$ and $\sigma_p^2 = 0.01$, both methods perform substantially worse, suggesting that this combination is unfavorable for the resulting posterior-based merge. Overall, the results show that the benefit of evidence-based alignment persists across a broad range of prior variances and likelihood temperatures. We therefore use the same setting as Talpini et al. (2025), with $\sigma_p^2 = 0.1$ and $T = 0.1$.

Table 7: Sensitivity to the prior variance σ_p^2 and the temperature T. Ten clients, random initialization, random local epochs, $\alpha = 0.1$. Test accuracy (%), mean ± SE over 5 seeds. Best of the two methods in bold.

σ_p^2	FMNIST $\Pi{=}I$	FMNIST FLAME (Our)	SVHN $\Pi{=}I$	SVHN FLAME (Our)	CIFAR-10 $\Pi{=}I$	CIFAR-10 FLAME (Our)
$T = 0.01$						
0.01	25.9±0.9	**29.1±2.1**	36.1±1.8	**46.1±3.4**	26.8±1.2	**29.6±2.1**
0.1	24.2±1.0	**27.8±3.6**	39.8±3.7	**51.6±5.0**	**26.8±1.3**	26.4±1.2
1.0	25.5±1.1	**26.6±2.7**	41.7±3.8	**53.0±4.0**	24.6±1.8	**24.8±1.2**
10.0	24.0±1.2	**25.7±3.2**	40.2±4.0	**50.2±4.9**	26.1±2.4	**27.5±1.3**
$T = 0.1$						
0.01	26.2±1.5	**36.2±1.2**	33.7±4.4	**46.4±4.5**	23.8±1.5	**25.4±1.6**
0.1	29.3±3.6	**38.6±2.4**	33.3±4.6	**54.2±3.8**	23.1±1.3	**27.3±2.1**
1.0	29.0±2.4	**35.5±2.7**	36.8±2.9	**49.5±2.2**	25.5±0.8	**26.8±1.7**
10.0	25.1±2.9	**29.5±2.5**	37.2±4.9	**50.4±5.1**	**25.6±1.4**	25.2±2.6
$T = 1.0$						
0.01	20.1±3.2	**22.8±2.4**	18.5±3.2	**19.5±3.2**	**13.5±1.1**	13.4±1.0
0.1	24.0±2.6	**30.7±2.1**	31.9±4.3	**51.8±3.8**	21.7±1.5	**27.7±2.0**
1.0	25.4±2.3	**27.3±2.6**	36.4±3.2	**42.7±3.6**	25.3±2.4	**28.2±3.0**
10.0	24.1±2.1	**32.2±1.7**	**40.2±5.7**	39.7±5.7	27.1±1.9	**28.0±2.9**

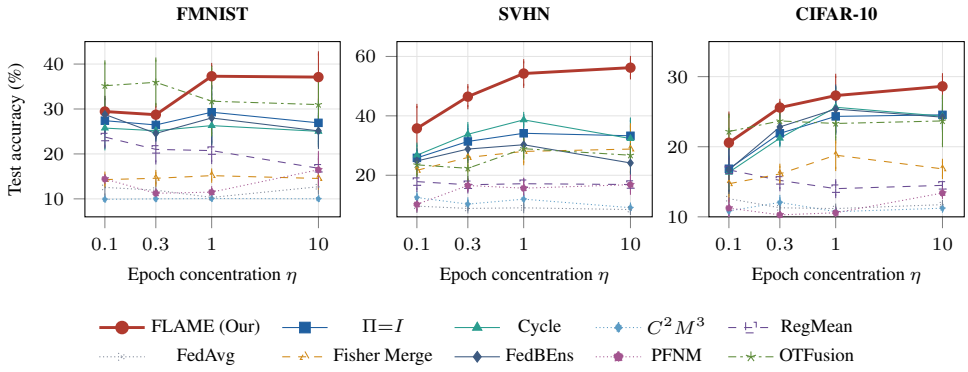

Figure 3: Test accuracy vs. epoch-distribution concentration η (small η = high variability). Ten clients, random init, heterogeniety α=0.1. Mean ± SE (5 seeds).

D.2 Epochs Heterogeneity

Figure 3 varies the concentration η of the Beta-Binomial epoch distribution (mean T_e, zero draws floored to one): large η approaches fixed epochs, whereas small η pushes clients to train either very briefly or for about $2T_e$ epochs. We use ten randomly initialized clients, $\alpha = 0.1$ and five seeds. Stronger epoch heterogeneity degrades all curvature-aware methods, since clients trained for few epochs have not reached a local optimum, where the Laplace approximation is meaningful, and

their posteriors are therefore less informative. FLAME remains the best method on SVHN for every η, with the largest margin at $\eta = 1$ (about 54% against 39% for Cycle and 34% for $\Pi{=}I$), and on CIFAR-10 for $\eta \geq 0.3$, where its advantage over $\Pi{=}I$ and Cycle is widest near fixed epochs ($\eta = 10$). On FashionMNIST, FLAME leads for $\eta \geq 1$, but OTFusion, which does not rely on curvature, is the strongest method at $\eta \leq 0.3$; OTFusion also slightly exceeds FLAME on CIFAR-10 at $\eta = 0.1$, within one standard error.