WHEN MULTI-DOMAIN FEDERATED LEARNING DOES NOT NEED DOMAIN-AWARE CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain heterogeneity in federated learning does not by itself require full domain-aware capacity. When domains differ only in class proportions, adjusting class biases can suffice. For a fixed representation and an observed domain context, we compare shared predictors that ignore context, prior-aware predictors that adjust only class biases, and full domain-aware predictors that may change the feature-to-label mapping across contexts. Under log-loss, we characterize the shared-to-domain Bayes gap through conditional mutual information and derive the exact residual after canonical prior correction. Combined with family- and protocol-dependent training-quality bounds, these gaps yield a sequential one-sided rule: certify a learned shared predictor when possible, otherwise test the prior-aware predictor, and abstain when neither certificate passes, without declaring richer capacity necessary. For full-participation local SGD on frozen affine heads, we derive a recoverability bound separating optimization transient, aggregated stochastic noise, local stochastic accumulation, and client-objective drift. Controlled studies audit the reference quantities and run the stopping rule on held-out calibration samples, using ground truth only to score its decisions. Real-data frozen-feature experiments show that full domain-aware heads can lower centralized log-loss, while ordinary shared aggregation often fails to recover this gain. Structure-aligned routing recovers substantially more but does not close the centralized gap. The framework separates the statistical value of domain-conditioned capacity from its federated recoverability.

1 INTRODUCTION

Federated learning (McMahan et al., 2017) trains models across decentralized sources that may differ in acquisition conditions, visual styles, and class composition (Li et al., 2025). A common response is to introduce local or domain-aware capacity, such as local normalization, domain-specific heads, and lightweight adaptation (Li et al., 2021; Zhong et al., 2023; Su et al., 2024; Lu et al., 2023). Yet domain heterogeneity alone does not establish the need for richer capacity, which may increase estimation difficulty and deployment costs. In centralized domain generalization, a related but distinct setting, well-tuned ERM baselines have also been found highly competitive with specialized algorithms (Gulrajani & Lopez-Paz, 2021; Teterwak et al., 2025). When domains differ only in class proportions, class-bias correction can suffice. We therefore ask: before introducing full domain-aware capacity, can a shared or lightly corrected predictor be certified sufficient?

Answering this question requires separating two issues: whether the context can improve prediction in the population, and whether a federated protocol can recover that improvement. We fix a representation $Z = f(X)$ and an observed context D, such as a recorded scanner type, supplied at training and prediction by a rule fixed before comparison. The practical choice is then whether to share the prediction rule (shared), adjust only class biases (prior-aware), or allow a context-specific feature-to-label mapping (full domain-aware). We study this capacity choice, rather than context discovery or generalization to unseen contexts.

Under log-loss, these levels yield two exact population gaps. The shared-to-domain Bayes gap $R^\star_{\mathrm{Sh}} - R^\star_{\mathrm{Dom}} = I(Y; D \mid Z)$ measures the label-relevant information remaining in the observed context after the representation is known. The exact prior-aware residual J_{PA} measures what remains relative to full domain-aware prediction after context-dependent prior correction. To inter-

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.

pret these gaps, we draw on the domain-adaptation taxonomy that distinguishes changes in class proportions from changes in class-conditional feature distributions, while allowing both to occur together (Zhang et al., 2013). In our analysis, the corresponding mechanisms are represented by prior coupling and class-conditional domain dependence. These mechanism quantities help interpret and upper-bound the exact gaps, which form the population components of the shared and prior-aware stopping tests.

Adding a valid family- and protocol-dependent training-quality bound to each gap yields a sequential one-sided decision. If the shared certificate does not exceed a prescribed tolerance ε, the learned shared predictor is ε-sufficient relative to every full domain-aware predictor. Otherwise, the procedure tests the learned prior-aware predictor and abstains if neither certificate passes. Population-level sufficiency does not automatically extend to a predictor returned by a federated protocol. The learned predictor must also be close to the relevant population reference, so federation enters through the protocol-dependent training-quality margin rather than the population gaps. For full-participation local SGD on frozen affine heads, we derive a bound separating the optimization transient, aggregated stochastic noise, local stochastic accumulation, and client-objective drift, and link the drift envelope to domain-level prior variation, within-context client label fragmentation, and class-conditional domain dependence. The contributions can be summarized as:

1. Under log-loss, we derive two exact population reference gaps: the shared-to-domain identity $R^\star_{\mathrm{Sh}} - R^\star_{\mathrm{Dom}} = I(Y; D \mid Z)$ and the residual J_{PA} of a canonical prior-aware reference relative to full domain-aware prediction. We decompose G_{cond} into J_{PA} and a marginal KL term, relate both gaps to prior coupling and class-conditional domain dependence, and derive a prototype upper-control route under a Gaussian shared-scale specialization.
2. We develop sequential one-sided certificates for learned shared and prior-aware predictors, with abstention when neither passes. Family- and protocol-dependent training-quality envelopes connect the population gaps to learned federated predictors. For affine heads on a fixed representation, we derive a four-component recoverability bound for full-participation local SGD with sample-size-weighted averaging, together with conditional guidance for local steps, client count and communication rounds.
3. We conduct theory-guided audits in controlled and real-data settings. Controlled studies audit reference quantities and mechanism distinctions, implement the stopping rule using held-out calibration samples, evaluate its decisions against known ground truth, and test protocol-component predictions. Frozen-feature affine heads and matched federated stress tests separate centralized capacity value from protocol-dependent recoverability.

2 Related Work

Federated capacity and optimization under heterogeneity. Federated methods either enrich the predictor or change its optimization. FedBN localizes batch normalization (Li et al., 2021), FedRep uses client-specific heads (Collins et al., 2021), and FedDAR uses domain-wise heads with a shared representation (Zhong et al., 2023). FedAPT and FedCLIP adapt prompts or lightweight modules (Su et al., 2024; Lu et al., 2023). Federated domain generalization instead targets unseen domains or clients through episodic learning, generalization adjustment, or cross-client style transfer (Liu et al., 2021; Zhang et al., 2023; Chen et al., 2023). These approaches attach parameters to clients or domains, target unseen domains, or condition on inferred sample features, whereas ours fixes an observed D available at prediction and asks whether shared or prior-aware capacity is sufficient. FedAvg, FedProx, and SCAFFOLD alter optimization (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020), while local-SGD analyses study update depth, stochastic noise, and heterogeneity (Khaled et al., 2020). We additionally ask whether the protocol recovers the simple family required by the certificate.

Shift mechanisms and prior correction. Domain adaptation distinguishes target shift (P_Y changes, $P_{X|Y}$ remains fixed), conditional shift (P_Y remains fixed, $P_{X|Y}$ changes under structural constraints), and generalized target shift, where both may change under related constraints (Zhang et al., 2013). For target shift, posterior-adjustment and black-box methods estimate new class priors and correct classifier outputs without relearning the class-conditional model (Saerens et al., 2002; Lipton et al., 2018). Two-sample diagnostics such as MMD (Gretton et al., 2012) compare distributions but do not quantify the residual label value of D given Z. These works characterize, estimate, or correct

shift. Related additive class-offset corrections appear in long-tail classification and federated training under client label skew (Menon et al., 2021; Zhang et al., 2022), but those works modify training or prediction for a fitted model rather than define a family-level sufficiency reference. Our analysis instead uses $I(Y; D \mid Z)$ and J_{PA} as exact log-loss reference gaps for shared and prior-aware sufficiency tests.

Model selection, risk control, and abstention. Structural risk minimization balances empirical error against hypothesis-class complexity (Shawe-Taylor et al., 1998). Risk-control methods use calibration data to test candidate predictors against prespecified risk targets (Angelopoulos et al., 2025), while selective prediction rejects individual inputs to trade coverage for risk (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019). Our certificates instead test family-level ε-sufficiency relative to every full domain-aware predictor and abstain when the evidence is inconclusive.

3 Capacity-Sufficiency Framework

3.1 Problem Setting and Capacity Hierarchy

Let $(X, Y, D, U) \sim \mathbb{P}$, where $X \in \mathcal{X}$ is an input and $Y \in [C] := \{1, \dots, C\}$ is its class label. The variable $D \in \mathcal{D}$, where $\mathcal{D}$ is finite, denotes the data source, acquisition condition, or deployment setting associated with the input. The variable $U \in [M] := \{1, \dots, M\}$ indexes the client from which the sample is drawn. For example, D may specify a scanner or acquisition protocol, a camera or sensor, a deployment site, or a benchmark domain label (photo, sketch, painting, clipart). For a sample with $D = d$, the value d is supplied whenever a prior-aware or full domain-aware predictor makes a prediction. For the trainable predictors considered here, the corresponding context values are also available with the training examples. The meaning of D and the rule that supplies it are fixed before the predictor families are compared and do not use the current sample's unknown label. Although D may be statistically associated with Y, for example because class prevalence differs across sites, it need not be a unique or causal "true domain." It is the domain-related information available to the prediction system, which we call the *observed domain context*.

For each client u, let P_u denote the conditional distribution of (X, Y, D) given $U = u$, and let $\omega_u := \mathbb{P}(U = u)$. Population risks use the marginal mixture $P := \sum_{u=1}^{M} \omega_u P_u$. U defines the federated partition but is not a predictor input; it may, but need not, coincide with D. For example, clients may be hospitals and contexts scanner models, allowing several domains per client and the same domain across clients.

Fix a representation $f : \mathcal{X} \to \mathbb{R}^p$ and set $Z = f(X)$. All predictor classes below are compared on this same representation. This isolates how the predictor uses the observed context D from how the representation itself is learned. We compare

$$\mathcal{H}_{\mathrm{Sh}} \longrightarrow \mathcal{H}_{\mathrm{PA}} \longrightarrow \mathcal{H}_{\mathrm{Dom}}, \tag{1}$$

where the arrows denote progressively less restricted use of D. The calligraphic symbols denote predictor classes, whereas the corresponding lowercase symbols denote individual predictors: $h_{\mathrm{Sh}} \in \mathcal{H}_{\mathrm{Sh}}$, $h_{\mathrm{PA}} \in \mathcal{H}_{\mathrm{PA}}$, and $h_{\mathrm{Dom}} \in \mathcal{H}_{\mathrm{Dom}}$. A shared predictor $h_{\mathrm{Sh}}(y \mid z)$ uses only the representation Z. A prior-aware predictor uses D only through an additive class-bias vector:

$$h_{\mathrm{PA}}(y \mid z, d) = \big[\mathrm{softmax}\big(s(z) + b_d\big)\big]_y, \tag{2}$$

where $s : \mathbb{R}^p \to \mathbb{R}^C$ is shared across all context values and $b_d \in \mathbb{R}^C$. The vector b_d can represent a context-dependent class-prior correction or, more generally, a class-specific logit offset. It does not make the feature-dependent score $s(z)$ itself context-specific. A full domain-aware predictor $h_{\mathrm{Dom}}(y \mid z, d)$ has no such additive restriction and may use a general score map $s_{\mathrm{Dom}} : \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^C$. These are capacity classes rather than fixed algorithms; concrete trainable instantiations are distinguished from the population comparators below.

Let Δ^{C-1} denote the probability simplex over C classes. For predictors from the three classes above, define the population log-loss risks

$$R_{\log}(h_{\mathrm{Sh}}) := \mathbb{E}_P\big[-\log h_{\mathrm{Sh}}(Y \mid Z)\big], \tag{3}$$

$$R_{\log}(h_A) := \mathbb{E}_P\big[-\log h_A(Y \mid Z, D)\big], \quad A \in \{\mathrm{PA}, \mathrm{Dom}\}. \tag{4}$$

The shared and full domain-aware Bayes risks, defined over all measurable predictors with the corresponding inputs, are

$$R^\star_{\mathrm{Sh}} := \inf_{h_{\mathrm{Sh}}:\mathbb{R}^p\to\Delta^{C-1}} R_{\log}(h_{\mathrm{Sh}}), \qquad R^\star_{\mathrm{Dom}} := \inf_{h_{\mathrm{Dom}}:\mathbb{R}^p\times\mathcal{D}\to\Delta^{C-1}} R_{\log}(h_{\mathrm{Dom}}). \tag{5}$$

For the prior-aware level, we instead use a canonical population reference, defined below, rather than a separate risk infimum over $\mathcal{H}_{\mathrm{PA}}$.

Given a training protocol Π, let $\widehat{h}_{\mathrm{Sh},\Pi}$ and $\widehat{h}_{\mathrm{PA},\Pi}$ denote the learned shared and prior-aware predictors. For $A \in \{\mathrm{Sh}, \mathrm{PA}\}$ and a prespecified log-loss tolerance $\varepsilon \geq 0$ (nats per example), we call $\widehat{h}_{A,\Pi}$ ε-sufficient relative to full domain-aware prediction if

$$R_{\log}(\widehat{h}_{A,\Pi}) - R^\star_{\mathrm{Dom}} \leq \varepsilon. \tag{6}$$

Since $R_{\log}(\widetilde{h}_{\mathrm{Dom}}) \geq R^\star_{\mathrm{Dom}}$ for every $\widetilde{h}_{\mathrm{Dom}} \in \mathcal{H}_{\mathrm{Dom}}$, Eq. (6) gives $R_{\log}(\widehat{h}_{A,\Pi}) - R_{\log}(\widetilde{h}_{\mathrm{Dom}}) \leq \varepsilon$.

3.2 Population Reference Gaps and Upper Controls

3.2.1 Shared-to-Domain Bayes Gap

Proposition 1 (Exact shared-to-domain Bayes gap)**.** *Under the population log-loss in Eqs.* (3) *and* (4)*, the Bayes risks in Eq.* (5) *are attained, up to null sets, by*

$$q^\star_{\mathrm{Sh}}(y \mid z) = p(y \mid z), \qquad q^\star_{\mathrm{Dom}}(y \mid z, d) = p(y \mid z, d). \tag{7}$$

Consequently, $$R^\star_{\mathrm{Sh}} = H(Y \mid Z), \qquad R^\star_{\mathrm{Dom}} = H(Y \mid Z, D), \tag{8}$$

and $$R^\star_{\mathrm{Sh}} - R^\star_{\mathrm{Dom}} = I(Y; D \mid Z). \tag{9}$$

This standard log-loss identity supplies the first reference gap in our hierarchy: $I(Y; D \mid Z)$ is the exact Bayes-level value of the observed context beyond the information already contained in Z.

Definition 2 (Prior coupling and class-conditional domain dependence)**.** The prior-coupling quantity is

$$G_{\mathrm{prior}} := I(Y; D) = \sum_{d\in\mathcal{D}} p(d)\ \mathrm{KL}(p(\cdot \mid d) \,\|\, p(\cdot))\,, \tag{10}$$

where the two distributions in the KL divergence are class-prior distributions. To express G_{cond} in KL form, define for each y with $p(y) > 0$ the pooled class-conditional law $p(z \mid y) := \sum_{d\in\mathcal{D}} p(d \mid y)\, p(z \mid y, d)$.

The class-conditional domain-dependence quantity is

$$G_{\mathrm{cond}} := I(Z; D \mid Y) = \sum_{y,d} p(y, d)\ \mathrm{KL}(p(z \mid y, d) \,\|\, p(z \mid y))\,, \tag{11}$$

where zero-mass terms are omitted.

The quantity G_{prior} measures how class priors vary with the observed context (e.g., disease prevalence may differ across hospitals), whereas G_{cond} measures whether the representation law remains context-dependent after the class is fixed (e.g., the representation distribution for the same diagnosis may differ across scanners or acquisition protocols). Accordingly, $G_{\mathrm{prior}} = 0$ if and only if $Y \perp D$, while $G_{\mathrm{cond}} = 0$ if and only if $Z \perp D \mid Y$. By the mutual-information chain rule (Appendix A.3),

$$I(Y; D \mid Z) = G_{\mathrm{prior}} + G_{\mathrm{cond}} - I(Z; D) \leq G_{\mathrm{prior}} + G_{\mathrm{cond}}. \tag{12}$$

Subtracting $I(Z; D)$, the context information already contained in the representation, leaves the label information that D adds once Z is observed. The sum $G_{\mathrm{prior}} + G_{\mathrm{cond}}$ is therefore only a conservative envelope: a large G_{cond} need not imply a large shared-to-domain gap when that context information is already encoded in Z.

3.2.2 PRIOR-AWARE-TO-DOMAIN REFERENCE GAP

Definition 3 (Canonical prior-aware population reference)**.** For almost every (z, d) under the population law of (Z, D), define

$$q^{\dagger}_{\mathrm{PA}}(y \mid z, d) := \frac{p(z \mid y)p(y \mid d)}{\sum_{y'} p(z \mid y')p(y' \mid d)}, \tag{13}$$

with zero-mass classes omitted. For classes with $p(y) > 0$, Bayes' rule also gives the equivalent proportional form $q^{\dagger}_{\mathrm{PA}}(y \mid z, d) \propto p(y \mid z)\, p(y \mid d)/p(y)$.

Relative to the full domain-aware Bayes rule, the first form replaces each context-specific class-conditional law $p(z \mid y, d)$ by the pooled law $p(z \mid y)$, while retaining the context-specific prior $p(y \mid d)$. The second form shows that, at the population-reference level, D enters through a class-specific prior-ratio correction. In log space, this correction is an additive class-specific offset, matching the structural restriction of the prior-aware family in Eq. (2). With this reference in place, define the prior-aware posterior residual

$$J_{\mathrm{PA}} := \sum_{d \in \mathcal{D}} p(d)\, \mathbb{E}_{p(z|d)} \left[\mathrm{KL}\Big(p(\cdot \mid z, d) \,\|\, q^{\dagger}_{\mathrm{PA}}(\cdot \mid z, d) \Big) \right], \tag{14}$$

where zero-mass contexts are omitted. Also define the reference representation marginal $q_d(z) := \sum_y p(y \mid d)p(z \mid y)$.

Proposition 4 (Exact prior-aware-to-domain reference gap)**.** *Under the population log-loss, the second population reference gap is*

$$R_{\log}(q^{\dagger}_{\mathrm{PA}}) - R^{\star}_{\mathrm{Dom}} = J_{\mathrm{PA}}. \tag{15}$$

$$\textit{Moreover,} \quad G_{\mathrm{cond}} = J_{\mathrm{PA}} + \sum_{d \in \mathcal{D}} p(d)\ \mathrm{KL}(p(z \mid d) \,\|\, q_d(z))\,. \tag{16}$$

$$\textit{Consequently,} \quad 0 \le J_{\mathrm{PA}} \le G_{\mathrm{cond}}. \tag{17}$$

Thus, J_{PA} is the canonical reference's excess log-loss over the full domain-aware Bayes rule, not necessarily the smallest gap achievable within the prior-aware family. The difference between G_{cond} and J_{PA} is exactly the marginal KL term in Eq. (16). Under pure prior shift ($Z \perp D \mid Y$), $q^{\dagger}_{\mathrm{PA}}(y \mid z, d) = p(y \mid z, d)$ almost surely under the population law of (Z, D), so $J_{\mathrm{PA}} = 0$ (Appendix A.6). This exactness motivates our canonical prior-aware reference. The shared-to-domain gap may nevertheless remain positive because the class prior $p(y \mid d)$ can still depend on d. Thus, the observed context can have Bayes-level value beyond a shared predictor even when the canonical prior-aware reference already attains the full domain-aware Bayes risk.

Appendix C defines the prototype dispersion $G^{\mathrm{proto}}_{\mathrm{cond}}$ (Eq. (205)), which summarizes squared deviations of context-specific class means from pooled class means, normalized by pooled class-covariance traces. Under the class-context Gaussian shared-scale specialization (Assumption 3), the appendix constructs a mean-shift envelope B^{τ}_{cond} using the within-context variance scale. Proposition 29 gives $G_{\mathrm{cond}} \le B^{\tau}_{\mathrm{cond}}$, hence $I(Y; D \mid Z) \le G_{\mathrm{prior}} + B^{\tau}_{\mathrm{cond}}$ and $J_{\mathrm{PA}} \le B^{\tau}_{\mathrm{cond}}$. The raw $G^{\mathrm{proto}}_{\mathrm{cond}}$ is an effect-size diagnostic, not a standalone numerical upper control for G_{cond}.

3.3 SEQUENTIAL ONE-SIDED CERTIFICATES

To transfer $I(Y; D \mid Z)$ and J_{PA} to predictors returned by protocol Π, we require valid upper envelopes on their learned-to-reference risk differences.

Assumption 1 (Valid training-quality envelopes)**.** For learned predictors $\widehat{h}_{\mathrm{Sh},\Pi}$ and $\widehat{h}_{\mathrm{PA},\Pi}$ returned by protocol Π, assume that there exist finite nonnegative quantities $\eta_{\mathrm{Sh}}(\Pi)$ and $\eta_{\mathrm{PA}}(\Pi)$ such that

$$R_{\log}\Big(\widehat{h}_{\mathrm{Sh},\Pi}\Big) \le R^{\star}_{\mathrm{Sh}} + \eta_{\mathrm{Sh}}(\Pi), \qquad R_{\log}\Big(\widehat{h}_{\mathrm{PA},\Pi}\Big) \le R_{\log}\Big(q^{\dagger}_{\mathrm{PA}}\Big) + \eta_{\mathrm{PA}}(\Pi). \tag{18}$$

Proposition 5 (One-sided family-sufficiency certificates)**.** *Under Assumption 1, define* $\mathcal{C}_{\mathrm{Sh}}(\Pi) := I(Y; D \mid Z) + \eta_{\mathrm{Sh}}(\Pi)$ *and* $\mathcal{C}_{\mathrm{PA}}(\Pi) := J_{\mathrm{PA}} + \eta_{\mathrm{PA}}(\Pi)$*. Then, for every full domain-aware comparator* $\widetilde{h}_{\mathrm{Dom}} \in \mathcal{H}_{\mathrm{Dom}}$*,*

$$R_{\log}\left(\widehat{h}_{\mathrm{Sh},\Pi}\right) - R_{\log}\left(\widetilde{h}_{\mathrm{Dom}}\right) \le \mathcal{C}_{\mathrm{Sh}}(\Pi), \qquad R_{\log}\left(\widehat{h}_{\mathrm{PA},\Pi}\right) - R_{\log}\left(\widetilde{h}_{\mathrm{Dom}}\right) \le \mathcal{C}_{\mathrm{PA}}(\Pi). \quad (19)$$

In particular, $R_{\log}(\widehat{h}_{A,\Pi}) - R^{\star}_{\mathrm{Dom}} \le \mathcal{C}_A(\Pi)$ *for* $A \in \{\mathrm{Sh}, \mathrm{PA}\}$. *Hence, for the tolerance* ε *in Eq.* (6)*,* $\mathcal{C}_A(\Pi) \le \varepsilon$ *certifies* $\widehat{h}_{A,\Pi}$ *as* ε*-sufficient.*

The certificates are evaluated in increasing order of capacity. We stop at $\widehat{h}_{\mathrm{Sh},\Pi}$ if $\mathcal{C}_{\mathrm{Sh}}(\Pi) \le \varepsilon$, otherwise at $\widehat{h}_{\mathrm{PA},\Pi}$ if $\mathcal{C}_{\mathrm{PA}}(\Pi) \le \varepsilon$, and otherwise abstain without ruling out full domain-aware capacity. Failure means only that the corresponding sufficiency claim is uncertified, not that the next capacity level is necessary. Conservative forms use $I(Y; D \mid Z) \le G_{\mathrm{prior}} + G_{\mathrm{cond}}$ and $J_{\mathrm{PA}} \le G_{\mathrm{cond}}$.

4 Federated Recoverability

The certificates in Proposition 5 require finite nonnegative bounds $\eta_A(\Pi)$ on the risk of each learned family relative to its population reference. For a fixed population distribution, federation affects these bounds, not the reference gaps $I(Y; D \mid Z)$ and J_{PA}. We decompose each bound as

$$\eta_A(\Pi) = \alpha_A^{\mathrm{approx}} + \epsilon_A^{\mathrm{opt}}(\Pi) + \epsilon_A^{\mathrm{gen}}(\Pi). \quad (20)$$

These terms cover approximation and regularization, federated optimization, and empirical-to-population transfer. The result below instantiates the first two, while the appendix gives a conditional finite-sample transfer.

Recoverability setting and conditions. Fix positive local sample sizes $n_1, \dots, n_M$, let $N := \sum_u n_u$ and $\omega_u := n_u/N$, and define $P^{(n)}(U = u) := \omega_u$ with $P^{(n)}(\cdot \mid U = u) := P_u$. All population quantities, objectives, and server weights below use this mixture, so the preceding certificates are instantiated for $P^{(n)}$. A different deployment mixture requires corresponding aggregation weights or an explicit mixture-shift allowance. We use affine frozen-feature heads

$$h_{\mathrm{Sh},\theta}(y \mid z) = \mathrm{softmax}(Wz + b)_y, \qquad h_{\mathrm{PA},\theta}(y \mid z, d) = \mathrm{softmax}(Wz + b + a_d)_y, \quad (21)$$

where θ collects W, b, and, for the prior-aware family, the offsets $\{a_d : d \in \mathcal{D}\}$, with ridge on all trainable coordinates. For $A \in \{\mathrm{Sh}, \mathrm{PA}\}$ and $\lambda_A > 0$, define

$$F^{\lambda}_{u,A}(\theta) := \mathbb{E}_{P_u}[-\log h_{A,\theta}(Y \mid Z, D)] + \frac{\lambda_A}{2}\|\theta\|_2^2, \qquad F^{\lambda}_A(\theta) := \sum_u \omega_u F^{\lambda}_{u,A}(\theta). \quad (22)$$

The shared head ignores D in this unified notation. Let θ^{λ}_A denote the minimizer of F^{λ}_A. Protocol Π runs full-participation local SGD for T communication rounds, using E local steps of size $\gamma > 0$ per round and sample-size-weighted averaging. Let $\theta_{A,t}$ denote the server iterate after round t, so the returned parameter is $\theta_{A,T}$. Assume $\|Z\|_2 \le B$ almost surely and conditionally unbiased stochastic gradients with bounded noise second moments. Let L_A be a common smoothness constant for the regularized client objectives, and take $\gamma L_A E \le 1/6$. The remaining stochastic-gradient conditions are in Appendix B.

Additional condition for the diagnostic bridge.

Assumption 2 (Client-refined factorization)**.** For every client u, $p_u(z, y, d) = p_u(d)p_u(y \mid d)p(z \mid y, d)$. Equivalently, $Z \perp U \mid (Y, D)$.

Thus $p(z \mid y, d)$ is shared across clients, while the client marginals $p_u(z)$ may differ through $p_u(d)$ and $p_u(y \mid d)$. Together with the appendix support and finite-KL conditions, this assumption is used only in Proposition 8. Theorem 6 and Corollary 7 retain an abstract heterogeneity envelope and do not require Assumption 2. Their reference-centered form uses the appendix finite-projection condition to define r_A.

To state the four components of the recoverability bound, write $R^{\mathrm{ref}}_{\mathrm{Sh}} := R^{\star}_{\mathrm{Sh}}$ and $R^{\mathrm{ref}}_{\mathrm{PA}} := R_{\log}(q^{\dagger}_{\mathrm{PA}})$ for the two population reference risks. Define the nonnegative regularized optimization gap at round t, $\Delta_{A,t} := F^{\lambda}_A(\theta_{A,t}) - F^{\lambda}_A(\theta^{\lambda}_A)$, and the ridge-adjusted allowance $\alpha^{\lambda}_A := [F^{\lambda}_A(\theta^{\lambda}_A) - R^{\mathrm{ref}}_A]_+$ between the regularized family optimum and its population reference, where $[x]_+ := \max\{x, 0\}$. If $\sigma^2_{u,A}$ uniformly bounds client u's conditional gradient-noise second moment, let

$$\sigma^2_{A,\mathrm{agg}} := \sum_u \omega_u^2 \sigma^2_{u,A}, \qquad \sigma^2_{A,\mathrm{loc}} := \sum_u \omega_u \sigma^2_{u,A}. \tag{23}$$

The ω_u^2 weights in $\sigma^2_{A,\mathrm{agg}}$ reflect server averaging, whereas $\sigma^2_{A,\mathrm{loc}}$ controls stochastic variation accumulated during local updates. The appendix fixes a minimum-norm reference projection $\bar{\theta}_A$, constructs r_A so that all local and server iterates and the population ridge minimizer lie within r_A of $\bar{\theta}_A$, and lets $\mathfrak{Z}_A(r_A)$ be a finite deterministic upper bound on the localized weighted client-gradient disagreement. The empirical extension also covers the empirical ridge minimizer.

Theorem 6 (Expected federated recoverability)**.** *Under the preceding protocol and stochastic-gradient conditions, for each $A \in \{\mathrm{Sh}, \mathrm{PA}\}$, there are explicit universal constants $c_0, c_1, c_2, c_3 > 0$, given in the appendix, such that with $\rho_A := 1 - c_0 \lambda_A \gamma E \in (0, 1)$,*

$$\begin{aligned} \mathcal{B}^{\mathrm{opt}}_A &:= \rho_A^T \Delta_{A,0} + c_1 \tfrac{L_A}{\lambda_A} \gamma \sigma^2_{A,\mathrm{agg}} + c_2 \tfrac{L_A^2}{\lambda_A} \gamma^2 (E-1) \sigma^2_{A,\mathrm{loc}} \\ &\quad + c_3 \tfrac{L_A^2}{\lambda_A} \gamma^2 (E-1)(2E-1) \mathfrak{Z}_A(r_A), \\ \mathbb{E}[\Delta_{A,T}] &\le \mathcal{B}^{\mathrm{opt}}_A, \ \mathbb{E}[R_{\log}(h_{A,\theta_{A,T}})] - R^{\mathrm{ref}}_A \le \alpha^{\lambda}_A + \mathcal{B}^{\mathrm{opt}}_A. \end{aligned} \tag{24}$$

Both expectations are over the local-SGD randomness.

At the expectation level, α^{λ}_A and $\mathcal{B}^{\mathrm{opt}}_A$ instantiate the approximation/reference and optimization components of Eq. (20), respectively. The four terms in $\mathcal{B}^{\mathrm{opt}}_A$ are the optimization transient, aggregated stochastic noise, local stochastic accumulation, and client-objective drift. The last two are distinct: local stochastic accumulation can remain with homogeneous client objectives, while client-objective drift can remain with deterministic local gradients.

Corollary 7 (High-probability recoverability bound)**.** *For each $A \in \{\mathrm{Sh}, \mathrm{PA}\}$ and $\delta_A \in (0, 1)$, with probability at least $1 - \delta_A$ over the local-SGD randomness,*

$$R_{\log}(h_{A,\theta_{A,T}}) \le R^{\mathrm{ref}}_A + \alpha^{\lambda}_A + \frac{\mathcal{B}^{\mathrm{opt}}_A}{\delta_A}. \tag{25}$$

On this event, $\alpha^{\lambda}_A + \mathcal{B}^{\mathrm{opt}}_A/\delta_A$ is a valid $\eta_A(\Pi)$, by applying Markov's inequality to the nonnegative $\Delta_{A,T}$ and then the risk-transfer bound in Appendix B. The $1/\delta_A$ conversion is conservative, and a union bound over the two branches gives simultaneous validity (Appendix B.6). No numerical tightness or real-data non-vacuity is claimed.

We now invoke Assumption 2 only to instantiate $\mathfrak{Z}_A$. Define the within-context client prior fragmentation $F^{\mathrm{FL}}_{\mathrm{prior}} := I(Y; U \mid D)$. Let β^{ref}_A denote the sample-size-weighted affine reference-projection KL residual defined in Appendix B.

Proposition 8 (Diagnostic-to-heterogeneity envelopes)**.** *Under the preceding sample-size-weighted population setup and the affine-head, boundedness, and smoothness conditions, Assumption 2, and the support, finite-KL, and finite-projection conditions in Appendix B, the localized heterogeneity envelopes in Theorem 6 may be chosen so that, for every $r \ge 0$,*

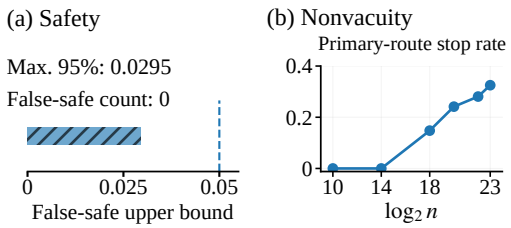
$$\mathfrak{Z}_{\mathrm{Sh}}(r) \lesssim \beta^{\mathrm{ref}}_{\mathrm{Sh}} + G_{\mathrm{prior}} + F^{\mathrm{FL}}_{\mathrm{prior}} + G_{\mathrm{cond}} + r^2, \qquad \mathfrak{Z}_{\mathrm{PA}}(r) \lesssim \beta^{\mathrm{ref}}_{\mathrm{PA}} + F^{\mathrm{FL}}_{\mathrm{prior}} + G_{\mathrm{cond}} + r^2. \tag{26}$$

Here $\lesssim$ hides only the explicit finite family-dependent geometry and loss-smoothness constants derived in the appendix.

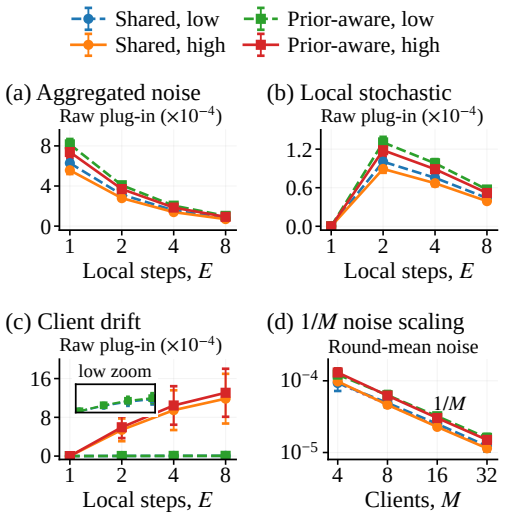
Thus G_{prior} is absent only from the prior-aware derived upper envelope, while $F^{\mathrm{FL}}_{\mathrm{prior}}$ and G_{cond} remain. This does not order the families' actual drifts or final risks, and G_{cond} cannot be replaced by J_{PA} without a separate drift lemma.

Table 1: Capacity value, federated recoverability, and controlled diagnostics. Left: stopping-rule calibration and full-participation component diagnostics. Right: full evaluation grids (Panel A) and 216 common matched cases, with full participation for every federated method (Panel B).

Stopping-rule calibration



Protocol-component diagnostics



Panel A. Broad-grid capacity value

Training setting	$\Delta_{\text{Sh}\to\text{PA}}$ Median (Positive)	$\Delta_{\text{PA}\to\text{Dom}}$ Median (Positive)
Centralized *n = 144*	+0.00348 (144/144)	+0.01235 (109/144)
Ordinary FedAvg *n = 11,520*	+0.01068 (10,217/11,520)	-0.02327 (280/11,520)

Panel B. Matched recoverability

n = 216 per row; federated q = 1

Method	$\Delta_{\text{PA}\to\text{Dom}}$ Median (Positive)	Routed − method Median paired (Routed better)
Centralized result	+0.01275 (176/216)	-0.01736 (73/216)
FedAvg	-0.04493 (0/216)	+0.05124 (216/216)
SCAFFOLD-style	-0.04349 (1/216)	+0.04755 (216/216)
FedProx, μ = 1	-0.04264 (0/216)	+0.04212 (216/216)
FedAdam	-0.06248 (18/216)	+0.07651 (192/216)
Routed WA†	-0.00264 (89/216)	reference —

Notes. Panels A–B. n counts evaluated cases; entries give medians with positive-case counts out of n in parentheses. Panel B's last column uses paired differences in $\Delta_{\text{PA}\to\text{Dom}}$ (Routed minus method), not differences of medians. Positive gains mean lower test log-loss for richer fitted affine heads, not unrestricted Bayes comparisons. M and K count total and participating clients; $q = K/M$ and real-data $M \in \{20, 40, 100\}$. Panel A is a broad-grid robustness summary, not a matched protocol comparison. FedAdam uses validation-selected $\eta_s = 2$ from a prespecified grid. † Fixed-Z, observed-domain-routed, head-only FedDAR-WA specialization. Dataset mappings, tuning, FedYogi robustness, and implementation differences are in the appendix.

Calibration (upper left). Here n is calibration sample size. Across 3,654 prespecified cells (100 replicates each), (a) reports no observed false-safe stops and a maximum cellwise one-sided 95% Clopper–Pearson upper bound of $0.0295 < 0.05$; (b) shows the primary-route stop rate rising with n, indicating increasing nonvacuity.

Protocol diagnostics (lower left). (a)–(c) use $M = K = 8$, $\gamma = \gamma_0/E$; (d) uses $M = K \in \{4, 8, 16, 32\}$, $E = 1$. Low/high denote controlled client-gradient disagreement, not component magnitude. Points are family–disagreement cell means; bars show one standard deviation over five seeds. All 20 matched trajectories (four series across five seeds) show these patterns. These trajectory-local diagnostics are neither an additive decomposition of final risk nor tests of exact theorem constants.

The same decomposition yields conditional guidance for E, M, and T. Under $\gamma = \gamma_0/E$ with $\gamma_0 L_A \le 1/6$, the coefficients of aggregated stochastic noise, local stochastic accumulation, and client-objective drift scale respectively as

$$\frac{1}{E}, \qquad \frac{E-1}{E^2}, \qquad \frac{(E-1)(2E-1)}{E^2}, \tag{27}$$

so changing E creates a three-component trade-off rather than a universal monotonic effect (audited in Table 1, lower left). If $\omega_u = 1/M$ and $\sigma^2_{u,A} \le \sigma^2_A$ for every client, then $\sigma^2_{A,\text{agg}} \le \sigma^2_A/M$, which

reduces only the aggregated stochastic noise component, and increasing T shrinks only the transient since $0 < \rho_A < 1$. The theorem is scoped to the population mixture and local-SGD protocol above. Other mixtures, partial participation, and alternative optimization or aggregation schemes require separate analysis, and the corresponding experiments are interpreted as mechanism studies.

5 Theory-Guided Experiments

We test three theory-aligned questions: capacity regimes distinguished by the population gaps, safe-side stopping on controlled distributions, and federated recovery of centralized gains from richer frozen-feature heads. Controlled audits use synthetic distributions; real-data experiments use PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), VLCS (Fang et al., 2013), and the common 100-class subset of DomainNet (Peng et al., 2019), with fixed, L2-normalized OpenCLIP ViT-B/16 features (Radford et al., 2021). We instantiate $\mathcal{H}_{\mathrm{Sh}}$, $\mathcal{H}_{\mathrm{PA}}$, and $\mathcal{H}_{\mathrm{Dom}}$ as shared, prior-aware (shared feature weights, domain-specific class biases), and per-domain affine heads, respectively. Define $\Delta_{A\to B} := \widehat{R}^{\mathrm{test}}_{\log}(\widehat{h}_A) - \widehat{R}^{\mathrm{test}}_{\log}(\widehat{h}_B)$; positive values favor B. These held-out fitted-head gains are not estimates of $I(Y; D \mid Z)$ or J_{PA}. Detailed settings are in the appendix.

The discrete constructions realize prior-only variation ($J_{\mathrm{PA}} = 0$), nuisance class-conditional variation ($G_{\mathrm{cond}} > 0$ but $I(Y; D \mid Z) = J_{\mathrm{PA}} = 0$), and semantic variation (both reference gaps are positive). Stopping decisions use independent calibration samples, with exact population quantities entering only the evaluation; both branches certify at nonzero rates (Table 3, Appendix D.2). The Stopping-rule calibration module of Table 1 shows safe-side stopping and increasing nonvacuity with more calibration data, while uncertified cases remain abstentions. Full discrete, calibration, and Gaussian prototype-chain results appear in Appendices D.1, D.2, and D.3, respectively.

The Protocol-component diagnostics module of Table 1 tests the E-scaling in Eq. (27) and the $1/M$ bound: aggregated stochastic noise falls with E, local stochastic accumulation is zero at $E = 1$ and peaks at $E = 2$, client-objective drift rises with E, and M-arm aggregated noise follows approximately $1/M$. These patterns support the component scalings in $\mathcal{B}_A^{\mathrm{opt}}$, which Corollary 7 uses to construct a valid $\eta_A(\Pi)$ for certificate $\mathcal{C}_A(\Pi)$, without implying monotonic final risk. Panels A and B of Table 1 separate centralized capacity value from federated recovery. The PA-to-Dom gain is often positive centrally but usually negative under ordinary FedAvg, while the Sh-to-PA control remains broadly positive, indicating a recoverability shortfall for the per-domain heads rather than a failed frozen-feature pipeline. On the matched full-participation cases, SCAFFOLD-style (Karimireddy et al., 2020), FedProx (Li et al., 2020), and validation-selected FedAdam (Reddi et al., 2021) retain shared aggregation and do not reverse this pattern. Routed WA instead aggregates separate observed-domain heads with domain-local sample-count weights and recovers substantially more value, though still less than centralized training. At $q = 1$, it is a frozen-feature, observed-domain-routed, head-only FedDAR-WA specialization (Zhong et al., 2023). The ordered ablation in E.2 attributes most of the routed gain to domain-local reweighting, so poor recovery under shared aggregation does not by itself show that domain-aware capacity is unnecessary; Appendices D.6 and E.1–E.4 give implementation details, the 24-cell breakdown, and partial-participation analyses.

6 Conclusion

Under log-loss, we derive the exact shared-to-domain gap and canonical prior-aware residual, combine them with training-quality envelopes in sequential one-sided certificates, and analyze the local-SGD training-quality margin. This separates the population value of domain-conditioned capacity from its federated recoverability. Controlled audits support the mechanism distinctions and safe-side stopping. On frozen features, richer capacity can have centralized value that the studied shared-aggregation protocols, including optimizer-corrected variants, often fail to recover; structure-aligned routing narrows but does not close the centralized gap. The rule certifies a learned shared or prior-aware predictor as ε-sufficient, or abstains without declaring richer capacity necessary. Future work includes relaxing the fixed representation and prespecified context through end-to-end representation learning, latent-context discovery, and prediction under uncertain or unavailable test-time context; extending the local-SGD envelopes to partial participation, other optimization and aggregation schemes, and richer families; and calibrated, nonvacuous real-data certificates bridging finite-dimensional affine-head risks to unrestricted Bayes comparators.

AI USE STATEMENT

We used large language models to assist with improving the clarity and readability of the manuscript, identifying relevant literature, developing and checking mathematical proofs, and writing experimental code, including code for generating synthetic datasets according to the data-generating procedures described in the paper. The authors reviewed and revised the AI-assisted material, checked the cited sources and mathematical arguments, and tested the code used in the experiments. The authors take full responsibility for the final content of this work, including the text, claims, proofs, code, and reported experimental results.

ETHICS STATEMENT

We identified no specific ethical concerns associated with this work. This work studies theoretical and algorithmic questions in federated learning using controlled synthetic distributions and existing benchmark image datasets. It does not involve new human-subject data collection or deployment in high-stakes settings.

REPRODUCIBILITY STATEMENT

Complete proofs and experimental details are provided in the appendices, including controlled-data constructions, dataset splits, client partitions, optimization settings, software and hardware configurations, and sources of randomness. We report the seeds or replication counts used in the main experiments. Upon publication, we plan to release the experiment and analysis code, environment specification, controlled-study configurations, DomainNet-common100 class list, split and client-partition manifests, protocol grids, and scripts required to reproduce the reported tables and figures.

REFERENCES

Mohammad A. Alhejji and Graeme Smith. A tight uniform continuity bound for equivocation. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 2270–2274, 2020.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

Junming Chen, Meirui Jiang, Qi Dou, and Qifeng Chen. Federated domain generalization for image recognition via cross-client style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 361–370, 2023.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/2331986`.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1657–1664, December 2013.

Mark Fannes. A continuity property of the entropy density for spin lattice systems. *Communications in Mathematical Physics*, 31:291–294, 1973.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local SGD on identical and heterogeneous data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4519–4529. PMLR, 2020.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6YEQUn0QICG.

Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, Ilir Murturi, Min Huang, and Schahram Dustdar. Federated domain generalization: A survey. *Proceedings of the IEEE*, 113 (4):370–410, 2025. doi: 10.1109/JPROC.2025.3596173.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021. doi: 10.1109/CVPR46437.2021.00107.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. FedCLIP: Fast generalization and personalization for CLIP in federated learning. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. URL https://openreview.net/forum?id=wgO-OK0_CQ.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LkFG3lB13U5.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998. doi: 10.1109/18.705570.

Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 15117–15125, 2024.

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan A. Plummer. ERM++: An improved baseline for domain generalization. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8514–8524, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdú, and Marcelo J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26311–26329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22p.html.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 819–827, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zhang13d.html.

Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3954–3963, 2023.

Aoxiao Zhong, Hao He, Zhaolin Ren, Na Li, and Quanzheng Li. FedDAR: Federated domain-aware representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6P9Y25Pljl6.