FEDR^{2}DG: RECONCILING ADVERSARIAL ROBUSTNESS AND DOMAIN GENERALIZATION WITH SELECTIVE DEFENSE IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated adversarial training becomes difficult when clients differ in both data domain and defense budget, especially when the deployment domain is unseen during training. We study selective federated robust domain generalization, where only a subset of source clients can afford full adversarial training and no target-domain data are available during training. We derive a held-out robust-risk bound that separates source-local fit, a federated mechanism gap, class-conditional source coverage, label-prior mismatch, and a conditional residual. Motivated by this bound, we propose Federated Reconciliation of Adversarial Robustness and Domain Generalization (FedR^{2}DG), which combines budget-aware local defense with drift control and class-conditional alignment. Experiments on PACS, OfficeHome, and mini-DomainNet50 show that FedR^{2}DG improves aggregate robust accuracy against AutoAttack by 4.91–11.99 percentage points over the best baseline in each dataset–budget setting. It also achieves the highest harmonic mean of standard and robust accuracy (H-score) among the compared methods, while largely preserving standard accuracy. The full-run computational cost per weakly defended client is 1.22–$1.43\times$ that of the matched selective-defense baseline. These results indicate that FedR^{2}DG improves the standard-robust balance under heterogeneous defense budgets without requiring full adversarial training at every client.

1 INTRODUCTION

Federated learning (FL) enables collaborative training without centralizing raw data (McMahan et al., 2017). Deployed models must also remain reliable under worst-case input perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). Although adversarial training (AT) provides a principled min–max framework for centralized robustness (Madry et al., 2018; Zhang et al., 2019; Yang et al., 2023), federated AT must additionally handle client drift, feature shift, heterogeneous local objectives, and model aggregation (Li et al., 2020; Karimireddy et al., 2020; Li et al., 2021; Guo et al., 2023; Chen et al., 2024).

Existing robust FL methods address distribution skew, selective defense, and federated min–max optimization (Chen et al., 2022; Hong et al., 2023; Li et al., 2023), but primarily study participating clients rather than robustness transfer to an *unseen* domain, where the clean distribution shifts and adversarial robustness must also generalize to unseen target samples (Gokhale et al., 2022; Zou & Liu, 2023).

We study selective federated robust domain generalization (**SFRDG**), illustrated in Figure 1. Source clients differ in both domain and defense budget. Only a subset can run full AT under heterogeneous resource constraints, and the global model is evaluated on clean and adversarial examples from a held-out domain without using target data or statistics during training. This setting couples two sources of difficulty. Heterogeneous defense budgets create uneven local robustness and can amplify federation-induced drift, while deployment on an unseen domain additionally requires robustness to transfer beyond the observed source domains. Strong adversarial fitting on the defended clients therefore does not by itself address aggregation-induced degradation or cross-domain robust transfer, motivating mechanisms that limit federated drift and preserve class-conditional structure across source domains.

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.

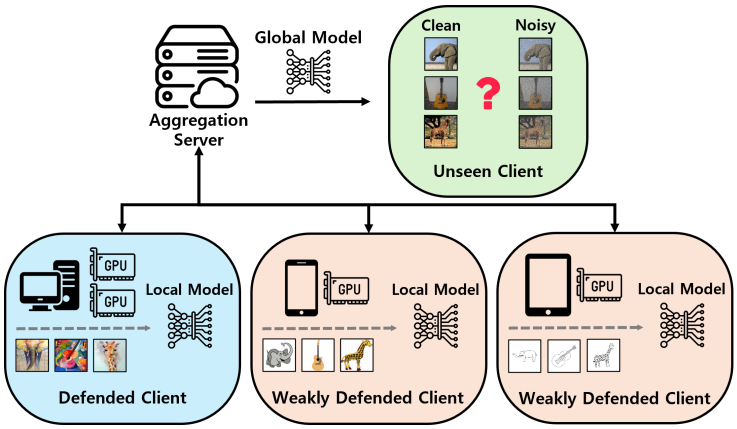


Figure 1: Selective federated robust domain generalization with heterogeneous source domains and defense budgets. Defended clients perform full AT, whereas weakly defended clients use budget-aware robust updates. The global model is evaluated on unseen clean and adversarial target inputs without using target-domain data during training.

Motivation and contributions. In a cross-silo federation, each institution can define a source domain, while a new institution may appear only at deployment. Domain generalization (DG) excludes target data from both training and model selection. Existing robust FL and federated DG leave open how selective defense and federated aggregation jointly affect robustness on unseen domains (Hong et al., 2023; Li et al., 2023; Zhang et al., 2023). We make three contributions. **(1)** We formulate SFRDG and derive a held-out robust-risk bound separating source-local fit, a federated mechanism gap, class-conditional source coverage, label-prior mismatch, and a conditional residual. We bound the residual by source conditional hull approximation and mixture-weight mismatch. **(2)** We propose Federated Reconciliation of Adversarial Robustness and Domain Generalization (FedR^{2}DG), which combines budget-aware local defense with drift control and class-conditional alignment. **(3)** Across three benchmarks and three perturbation budgets, FedR^{2}DG consistently improves held-out-domain robust accuracy and the standard–robust balance, while the total cost per weakly defended client remains only 1.22–$1.43\times$ that of the matched selective-defense baseline.

2 Related Work

We review prior work on federated learning, robustness, and domain generalization.

Federated Learning under Heterogeneity. FedProx (Li et al., 2020) addresses statistical and systems heterogeneity through proximal regularization and variable local work, while SCAFFOLD (Karimireddy et al., 2020) corrects client drift using control variates. FedBN (Li et al., 2021) uses local batch normalization to address feature shift. Domain-skewed FL further studies local consistency and domain diversity to improve fairness across heterogeneous domains (Chen et al., 2024).

Adversarial Robustness in Federated Learning. PGD-AT (Madry et al., 2018) provides a robust optimization baseline. CalFAT (Chen et al., 2022) addresses label skew in federated adversarial training. Federated Robustness Propagation (FRP) transfers robustness from resource-rich to low-resource clients (Hong et al., 2023). FAL (Li et al., 2023) analyzes convergence of federated adversarial learning. These methods primarily address robustness among participating clients; FedR^{2}DG instead targets robustness transfer to a held-out domain under selective defense.

Domain Generalization in Federated Learning. DG learns from labeled sources without target-domain access. DICA, IRM, and REx pursue invariant representations or predictors and stable source risks (Muandet et al., 2013; Arjovsky et al., 2019; Krueger et al., 2021), while DomainBed emphasizes reproducible evaluation and model selection (Gulrajani & Lopez-Paz, 2021). In federated settings, FedDG uses frequency-space episodic learning (Liu et al., 2021), and FedDG-GA adjusts aggregation

for generalization (Zhang et al., 2023). FedProto shares class prototypes across heterogeneous clients (Tan et al., 2022). Our coverage condition connects to multi-source distributional geometry (Cortes et al., 2021). Our residual analysis bounds the gap for held-out class-conditionals in or near the source conditional convex hull by hull-approximation error and mixture-weight mismatch. Because representation alignment can fail under structured shifts or extrapolation (Mahajan et al., 2021; Rosenfeld et al., 2021), the analysis retains an explicit conditional residual.

Robustness under Domain Shift. Adversarial robustness and out-of-domain generalization are related but distinct objectives: improvements in domain generalization do not necessarily imply adversarial robustness, and models designed for OOD generalization can remain vulnerable to adversarial perturbations (Gokhale et al., 2022; Zou & Liu, 2023). TAROT studies robust transfer through target-risk analysis and unlabeled target data (Yang et al., 2025). FedR^{2}DG instead considers a source-only federated setting in which robustness must transfer to an unseen domain while only a subset of clients can afford full adversarial training.

3 Problem Setup and Robust-Risk Analysis

We formulate SFRDG and bound held-out robust risk and its conditional residual.

Setting and notation. Let $[K] = \{1, \dots, K\}$ index the domains and $[C]$ the classes. For held-out domain j, the source clients are $S_j = [K] \setminus \{j\} = \mathcal{D}_j \cup \mathcal{N}_j$, where $\mathcal{D}_j$ and $\mathcal{N}_j$ denote defended and weakly defended clients. Client i has distribution $\mathcal{P}_i$ over $(\mathbf{X}, Y)$ and sample $\widehat{D}_i = \{(\boldsymbol{x}_{i,m}, y_{i,m})\}_{m=1}^{n_i}$, with $n_{i,c}$ examples of class c. Positive weights ν_i sum to one and define the source mixture $\mathcal{P}_\nu = \sum_{i \in S_j} \nu_i \mathcal{P}_i$. Parenthesized superscripts index communication rounds and, when present, local optimization steps. At round T, client i returns the prediction state $\boldsymbol{\omega}_i^{(T)} = (\boldsymbol{\theta}_i^{(T)}, \boldsymbol{b}_i^{(T)})$, where $\boldsymbol{\theta}_i^{(T)}$ denotes trainable parameters and $\boldsymbol{b}_i^{(T)}$ collects non-trainable model buffers used for inference, including batch-normalization running means and variances. This state induces predictor $h_i^{(T)}$. The server forms $\bar{\boldsymbol{\omega}}^{(T)} = \sum_{i \in S_j} \nu_i \boldsymbol{\omega}_i^{(T)}$, which induces the aggregated predictor $\bar{h}^{(T)}$. Source laws, sample sizes, attack radius ε, temperature τ_{con}, and model classes are fixed background parameters. Function arguments expose dependence on learned models, representations, anchors, and mixture weights; j marks dependence on the target distribution. Table 3 summarizes the notation.

For perturbation radius ε, define the robust loss and risk by

$$g_h(\boldsymbol{x}, y) = \sup_{\|\boldsymbol{\delta}\|_\infty \le \varepsilon} \ell(h(\boldsymbol{x} + \boldsymbol{\delta}), y), \qquad \mathcal{R}_i^\varepsilon(h) = \mathbb{E}_{\mathcal{P}_i}[g_h(\mathbf{X}, Y)].$$

Perturbed inputs remain in the valid input domain. The analysis assumes a bounded loss $0 \le g_h \le B$, for example clipped cross-entropy; training uses ordinary cross-entropy. Thus the theorem concerns a bounded risk, not unbounded training cross-entropy. Write $\boldsymbol{\phi}$ for the feature extractor of $\bar{h}^{(T)}$ and $\mathbf{Z}_{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})/\|\boldsymbol{\phi}(\boldsymbol{x})\|_2$ for its normalized feature, assuming nonzero feature norms. Global class anchors $\mathbf{v} = (v_1, \dots, v_C)$ have unit norm. Their prototype loss is

$$\ell_{\mathrm{proto}}(\boldsymbol{x}, c; \boldsymbol{\phi}, \mathbf{v}) = -\log \frac{\exp(\langle \mathbf{Z}_{\boldsymbol{\phi}}(\boldsymbol{x}), v_c \rangle / \tau_{\mathrm{con}})}{\sum_{c'=1}^{C} \exp(\langle \mathbf{Z}_{\boldsymbol{\phi}}(\boldsymbol{x}), v_{c'} \rangle / \tau_{\mathrm{con}})},$$

where $\tau_{\mathrm{con}} > 0$ is the temperature.

Source fit and alignment. For a collection of local predictors $\mathbf{h} = (h_i)_{i \in S_j}$, define

$$\begin{aligned} \widehat{L}_{\mathrm{src}}(\mathbf{h}; \nu) &= \sum_{i \in S_j} \frac{\nu_i}{n_i} \sum_{m=1}^{n_i} g_{h_i}(\boldsymbol{x}_{i,m}, y_{i,m}), \\ \widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu) &= \sum_{i \in S_j} \frac{\nu_i}{n_i} \sum_{m=1}^{n_i} \ell_{\mathrm{proto}}(\boldsymbol{x}_{i,m}, y_{i,m}; \boldsymbol{\phi}, \mathbf{v}). \end{aligned} \tag{1}$$

Source fit depends on local predictors and mixture weights; alignment depends on the shared representation, anchors, and weights. Both use source samples. Source fit uses the true adversarial supremum: PGD and replay are training proxies, not upper bounds for it.

Aggregation drift. Write $\boldsymbol{\Delta}^{(T)} = (\Delta_i^{(T)})_{i \in S_j}$ for the client drift vector:

$$\Delta_i^{(T)} = \|\boldsymbol{\omega}_i^{(T)} - \bar{\boldsymbol{\omega}}^{(T-1)}\|_2, \qquad U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)}; \nu) = 2L_{\mathrm{stab}} M_\varepsilon \sum_{i \in S_j} \nu_i \Delta_i^{(T)}. \tag{2}$$

Thus the penalty depends on returned and broadcast model states and the aggregation weights. The constants L_{stab} and M_ε convert state drift into a robust-risk gap.

Label-prior mismatch and conditional residual. Let $\pi_i(c) = \mathcal{P}_i(Y = c)$ and $\pi_\nu = \sum_i \nu_i \pi_i$, with empirical versions $\widehat{\pi}_i(c) = n_{i,c}/n_i$ and $\widehat{\pi}_\nu = \sum_i \nu_i \widehat{\pi}_i$. The source prior radius $\widehat{r}_Y(\nu) = \max_i \|\widehat{\pi}_i - \widehat{\pi}_\nu\|_1$ depends on source class counts and weights, but not on the learned representation. For $\mathcal{C}_\nu = \{c : \pi_\nu(c) > 0\}$ and $r \in \{j, \nu\}$, define $q_{r,c}^h(z; \boldsymbol{\phi}) = \mathbb{E}_{\mathcal{P}_r}[g_h(\mathbf{X}, Y) \mid \mathbf{Z}_{\boldsymbol{\phi}}(\mathbf{X}) = z, Y = c]$. The remaining target–source gap is

$$\rho_j(h; \boldsymbol{\phi}, \nu) = \left| \sum_{c \in \mathcal{C}_\nu} \pi_j(c) \mathbb{E}_{\mathcal{P}_{j,\mathbf{Z}|Y=c}}[q_{j,c}^h(\mathbf{Z}; \boldsymbol{\phi}) - q_{\nu,c}^h(\mathbf{Z}; \boldsymbol{\phi})] \right|.$$

It depends jointly on the predictor h, conditioning representation $\boldsymbol{\phi}$, source mixture ν, and target law $\mathcal{P}_j$. It measures robust-loss differences at the same feature and label, which source data alone cannot identify. Conditional profiles use fixed measurable versions as specified in Appendix A.2; terms with $\pi_j(c) = 0$ are interpreted as zero.

Conditions and calibration. The assumptions in Appendix A.2 require source coverage of target feature and label-prior shifts, compatibility with a Lipschitz discrepancy class, and risk stability under prediction-state changes. The coefficient $\Lambda_{\mathrm{cov}}^j(\boldsymbol{\phi}, \mathbf{v}; \nu)$ combines coverage, discrepancy compatibility, and source-centroid dispersion; it depends on the representation, anchors, weights, and target law. The prior-coverage factor $\kappa_\pi^j(\nu)$ depends on the weights and target law, independently of $\boldsymbol{\phi}$ and $\mathbf{v}$. For $C \geq 2$, $\tau = \tau_{\mathrm{con}}$, and $u \geq b_{\tau,C} := \log(1 + (C-1)e^{-2/\tau_{\mathrm{con}}})$, set

$$\Psi_{\tau,C}(u) = \min\left\{2,\ 2 + \tau_{\mathrm{con}} \log \frac{e^u - 1}{C - 1}\right\}.$$

This nondecreasing, concave map converts prototype loss into angular deviation.

Finite-sample corrections. For $\beta \in (0, 1/3)$, the corrections $\mathcal{E}_{\mathrm{src}}(\beta; \nu)$ and $\mathcal{E}_{\mathrm{proto}}(\beta; \nu)$ bound source- and prototype-loss estimation errors using Rademacher complexity and concentration. The label-prior error bound $\mathcal{E}_{\mathrm{prior}}(\beta)$ uses concentration alone and is independent of ν. For i.i.d. samples within each client, fixed function classes, and fixed sampling-design weights, these corrections hold jointly and uniformly on an event $\mathcal{A}_\beta$ with probability at least $1 - 3\beta$. Appendix A.3 gives their exact definitions and derivation.

Theorem 1. *Fix j, T and $\beta \in (0, 1/3)$. Under the bounded-loss, coverage, discrepancy, and stability conditions in Appendix A.2, on $\mathcal{A}_\beta$,*

$$\begin{aligned} \mathcal{R}_j^\varepsilon(\bar{h}^{(T)}) \leq\ & \widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)}; \nu) + \mathcal{E}_{\mathrm{src}}(\beta; \nu) + U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)}; \nu) \\ & + \Lambda_{\mathrm{cov}}^j(\boldsymbol{\phi}, \mathbf{v}; \nu) \sqrt{\Psi_{\tau,C}\big(\widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu) + \mathcal{E}_{\mathrm{proto}}(\beta; \nu)\big)} \\ & + B\kappa_\pi^j(\nu)\big(\widehat{r}_Y(\nu) + \mathcal{E}_{\mathrm{prior}}(\beta)\big) + \rho_j(\bar{h}^{(T)}; \boldsymbol{\phi}, \nu). \end{aligned} \tag{3}$$

Remark 1. The decomposition motivates robust local updates, drift control, and prototype alignment (proof in Appendix A.4). Coverage, discrepancy compatibility, and stability are assumptions, and a smaller prototype loss can be offset by a larger $\Lambda_{\mathrm{cov}}^j(\boldsymbol{\phi}, \mathbf{v}; \nu)$ or residual. Training uses evolving local features, making its prototype loss a proxy for the bound's global-representation loss. Reweighting changes $\widehat{r}_Y(\nu)$ without guaranteeing a decrease. EMA directly contracts only the trainable-parameter contribution to U_{drift}; prediction-buffer drift remains in the bound.

Controlling the conditional residual. Write $\mathcal{P}_{r,c} := \mathcal{P}_{r,\mathbf{X}|Y=c}$ for the input class-conditional law whenever $\pi_r(c) > 0$, and use the convention $d_{\mathrm{TV}}(P, Q) := \sup_A |P(A) - Q(A)|$ over measurable sets A.

Lemma 1. *Suppose $0 \le g_h \le B$ and the fixed measurable source profiles $q^h_{\nu,c}(\,\cdot\,;\boldsymbol{\phi})$ take values in $[0, B]$, including on source-null sets. Then*

$$\rho_j(h;\boldsymbol{\phi},\nu) \le 2B \sum_{c\in\mathcal{C}_\nu} \pi_j(c) d_{\mathrm{TV}}(\mathcal{P}_{j,c}, \mathcal{P}_{\nu,c}). \tag{4}$$

For $c \in \mathcal{C}_\nu$, let $S_j(c) = \{i \in S_j : \pi_i(c) > 0\}$. Conditioning the source mixture gives

$$\mathcal{P}_{\nu,c} = \sum_{i\in S_j(c)} \nu_{i,c}\mathcal{P}_{i,c}, \qquad \nu_{i,c} = \frac{\nu_i \pi_i(c)}{\pi_\nu(c)}.$$

For weights $\alpha_c \in \Delta(S_j(c))$ in the probability simplex, set $Q_{\alpha,c} = \sum_{i\in S_j(c)} \alpha_{i,c}\mathcal{P}_{i,c}$ and $\nu_c = (\nu_{i,c})_{i\in S_j(c)}$.

Corollary 1. *Under the conditions of Lemma 1, for any such α_c,*

$$\rho_j(h;\boldsymbol{\phi},\nu) \le 2B \sum_{c\in\mathcal{C}_\nu} \pi_j(c) d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c}) + B \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\|\alpha_c - \nu_c\|_1. \tag{5}$$

If $\mathcal{P}_{j,c} = Q_{\alpha,c}$ for every source-supported class with positive target mass, only the mixture-mismatch term remains. If also $\alpha_c = \nu_c$, then $\rho_j(h;\boldsymbol{\phi},\nu) = 0$.

Remark 2. For each $c \in \mathcal{C}_\nu$ with $\pi_j(c) > 0$, the *source-hull approximation error*, $d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c})$, measures how well a mixture of source class-conditional input distributions with weights α_c approximates the target conditional $\mathcal{P}_{j,c}$. Minimizing over α_c gives the distance to the source conditional hull—the set of all such mixtures. The *mixture-weight mismatch*, $\|\alpha_c - \nu_c\|_1$, compares α_c with the weights induced by aggregation and source class priors, controlling the discrepancy between $Q_{\alpha,c}$ and $\mathcal{P}_{\nu,c}$. It concerns source proportions within a class, not class frequencies. Thus, accurate target approximation by source mixtures with weights close to ν_c implies a small residual. Exact hull membership alone does not guarantee a zero residual; matching $\mathcal{P}_{j,c} = \mathcal{P}_{\nu,c}$ for every contributing class is sufficient. These quantities are not estimated from target data during training or model selection. This reflects a general limitation of domain generalization: source data alone cannot guarantee low risk on arbitrary unseen domains without assumptions relating source and target distributions (Blanchard et al., 2021; Jalaldoust et al., 2024; Wang et al., 2026).

4 FedR^{2}DG

We describe global centroids, selective local updates, and model aggregation. All clients take E local optimizer steps per round; their defense computation differs. After T_{warm} rounds of clean FedAvg, the server initializes global centroids and runs T selective-defense rounds. Appendix B gives the complete procedure and component-to-bound mapping.

4.1 Global Centroids

For a class present on client i, the local and global centroids are

$$v_{i,c} = \mathrm{Normalize}\left(\frac{1}{n_{i,c}} \sum_{m:y_{i,m}=c} \mathbf{Z}_{\boldsymbol{\theta}_i}(\boldsymbol{x}_{i,m})\right), \qquad v_c = \mathrm{Normalize}\left(\sum_{i:n_{i,c}>0} p_{i,c} v_{i,c}\right). \tag{6}$$

Here $p_{i,c} = n_{i,c}/\sum_{i'\in S_j} n_{i',c}$ is the class-count weight, $\mathbf{Z}_{\boldsymbol{\theta}}$ is the normalized feature of $h_{\boldsymbol{\theta}}$ and $\mathrm{Normalize}(a) = a/\|a\|_2$ for nonzero a. We assume all normalization inputs are nonzero; absent client classes are omitted. All benchmark classes are represented in the source federation. Initialization uses the shared post-warm-up model. Every I_{cent} main rounds, clients recompute centroids in evaluation mode without gradients using their updated local models. The server forms $\widehat{v}_c = \mathrm{Normalize}(\sum_i p_{i,c} v_{i,c})$ and updates

$$v_c \leftarrow \mathrm{Normalize}\big(\alpha_{\mathrm{ema}} v_c + (1-\alpha_{\mathrm{ema}})\widehat{v}_c\big). \tag{7}$$

The class-count weights $p_{i,c}$ estimate class anchors, whereas the model weights ν_i below balance clients. The coefficient α_{ema} controls temporal smoothing of centroids and is distinct from the parameter-tether coefficient γ_{ema}.

4.2 Selective Local Updates

We use the prototype loss ℓ_{proto} defined above, with global centroids $\mathbf{v}$ as class anchors; it is evaluated only on clean features. At round t, each client initializes its trainable parameters and prediction buffers from the broadcast state $(\bar{\boldsymbol{\theta}}^{(t-1)}, \bar{\boldsymbol{b}}^{(t-1)})$. We write only the trainable-parameter update explicitly below. After every optimizer step, it interpolates its trainable parameters toward this fixed broadcast reference:

$$\boldsymbol{\theta}_i^{(t,k+1)} = (1-\gamma_{\text{ema}})(\boldsymbol{\theta}_i^{(t,k)} - \eta_i g_i^{(t,k)}) + \gamma_{\text{ema}}\bar{\boldsymbol{\theta}}^{(t-1)}, \qquad k = 0, \dots, E-1. \tag{8}$$

Here η_i is the learning rate and $g_i^{(t,k)}$ is the effective optimizer direction, including momentum and weight decay when used. For $\|g_i^{(t,k)}\|_2 \le G_i$ and $0 < \gamma_{\text{ema}} \le 1$,

$$\|\boldsymbol{\theta}_i^{(t)} - \bar{\boldsymbol{\theta}}^{(t-1)}\|_2 \le \eta_i G_i \frac{(1-\gamma_{\text{ema}})[1-(1-\gamma_{\text{ema}})^E]}{\gamma_{\text{ema}}}.$$

This replaces the untethered factor E by a discounted sum. It controls the parameter contribution to $U_{\text{drift}}(\boldsymbol{\Delta}^{(T)}; \nu)$; untethered prediction buffers remain part of the state drift in Equation (2).

Defended clients. For each mini-batch, client $i \in \mathcal{D}_j$ generates $\boldsymbol{x}_{\text{adv}}$ by PGD on cross-entropy with ℓ_∞ radius ε, then minimizes

$$\mathcal{L}_{\text{def}}(\boldsymbol{\theta}; \boldsymbol{x}, y, \mathbf{v}) = \text{CE}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_{\text{adv}}), y) + \lambda \ell_{\text{proto}}(\boldsymbol{x}, y; \boldsymbol{\phi}_{\boldsymbol{\theta}}, \mathbf{v}). \tag{9}$$

The prototype term always uses clean features. This supplies a clean alignment signal alongside the adversarial fitting objective.

Weakly defended clients. Client $i \in \mathcal{N}_j$ performs B_C clean steps followed by B_H mini-batches replayed ζ times each, with $B_C + B_H\zeta = E$. Clean steps minimize

$$\mathcal{L}_{\text{clean}}(\boldsymbol{\theta}; \boldsymbol{x}, y, \mathbf{v}) = \text{CE}(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y) + \lambda \ell_{\text{proto}}(\boldsymbol{x}, y; \boldsymbol{\phi}_{\boldsymbol{\theta}}, \mathbf{v}). \tag{10}$$

For each replay mini-batch, reset $\boldsymbol{\delta}^{(0)} = 0$. At replay r, use $\boldsymbol{x}^{(r-1)} = \text{clip}_{[0,1]}(\boldsymbol{x} + \boldsymbol{\delta}^{(r-1)})$ and the loss $\mathcal{L}_{\text{rep}} = \text{CE}(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(r-1)}), y)$. Update

$$\boldsymbol{\delta}^{(r)} = \text{clip}_{[-\varepsilon,\varepsilon]}\big(\boldsymbol{\delta}^{(r-1)} + \varepsilon\, \text{sign}(\nabla_{\boldsymbol{x}^{(r-1)}} \mathcal{L}_{\text{rep}})\big). \tag{11}$$

The same backward pass supplies the parameter gradient for the optimizer update and the input gradient for Equation (11). Following Free-AT (Shafahi et al., 2019), replays share computation, but our perturbations reset for each mini-batch and round. With zero initialization, the first replay is clean; subsequent replays use accumulated perturbations. We omit the prototype term in replay because the clean phase supplies it. Every clean or replay optimizer update is followed by Equation (8).

4.3 Server Aggregation

The server averages returned parameters and prediction buffers with

$$\nu_i = \frac{\sqrt{n_i}}{\sum_{i' \in S_j} \sqrt{n_{i'}}}, \qquad \bar{\boldsymbol{\theta}}^{(t)} = \sum_{i \in S_j} \nu_i \boldsymbol{\theta}_i^{(t)}, \qquad \bar{\boldsymbol{b}}^{(t)} = \sum_{i \in S_j} \nu_i \boldsymbol{b}_i^{(t)}. \tag{12}$$

These weights are the normalized geometric mean of data-proportional and uniform weights. They reduce the relative influence of data-rich clients while retaining sample-size information. Unlike centroid weights $p_{i,c}$, model weights are not class-specific. Their effect on the source mixture and label-prior radius motivates rebalancing, but the bound does not establish optimality of this choice.

5 Experiments

We evaluate held-out performance, hyperparameter sensitivity, and training cost.

Table 1: Held-out-domain standard accuracy, AA robust accuracy, and their harmonic mean (H-score, %). Results are averaged over target and defended-client choices, with H-score computed per case.

Method	ε	PACS			OfficeHome			mini-DomainNet50		
		Standard.	Rob.	H-score	Standard.	Rob.	H-score	Standard.	Rob.	H-score
Robust ERM	$2/255$	63.93	49.24	55.07	56.06	41.33	47.38	57.65	41.04	47.78
	$4/255$	62.99	40.14	47.44	53.63	32.42	39.99	53.63	30.78	38.73
	$8/255$	60.91	29.71	36.68	48.31	21.77	29.43	46.49	19.45	26.80
SR-FedAvg	$2/255$	85.25	66.19	73.98	68.77	46.98	55.45	73.86	49.03	58.59
	$4/255$	83.92	50.74	60.93	67.28	34.86	45.16	71.71	34.14	45.30
	$8/255$	80.83	32.45	41.32	63.94	20.76	30.25	67.03	18.88	27.95
SR-FRP	$2/255$	84.89	66.95	74.29	**68.98**	42.69	52.13	**74.61**	44.86	55.40
	$4/255$	**84.21**	47.57	57.96	**68.39**	22.96	33.33	**73.79**	24.23	35.00
	$8/255$	**82.61**	24.22	31.22	**67.01**	6.19	10.84	**72.17**	7.53	12.89
SR-FedProx	$2/255$	85.27	66.20	73.99	68.75	47.04	55.49	73.85	49.04	58.59
	$4/255$	84.00	50.81	61.01	67.25	34.86	45.13	71.73	34.03	45.18
	$8/255$	80.79	32.50	41.38	63.99	20.88	30.37	66.82	18.80	27.94
FedR^{2}DG	$2/255$	**85.41**	**72.64**	**78.34**	68.27	**51.95**	**58.80**	73.40	**55.23**	**62.88**
	$4/255$	83.94	**61.38**	**70.12**	66.89	**41.82**	**50.99**	71.58	**42.85**	**53.12**
	$8/255$	80.90	**44.49**	**54.64**	64.06	**28.66**	**38.84**	67.74	**27.34**	**37.73**

5.1 Experimental Setup

Datasets and Federated Configuration. We evaluate three domain generalization benchmarks under the leave-one-domain-out (LODO) protocol. For each held-out target domain, the remaining three domains serve as source clients: one defended client and two weakly defended clients (Section 4.2). Evaluating every target and defended source yields 12 cases per dataset. We select these benchmarks for their standard use in the domain generalization literature and for spanning a range of class counts and domain-shift severity. **PACS** (Li et al., 2017) contains four domains: Photo, Art Painting, Cartoon, and Sketch, with 7 classes. **OfficeHome** (Venkateswara et al., 2017) contains four domains: Art, Clipart, Product, and Real-World, with 65 classes. **mini-DomainNet50** is a 50-class subset constructed from the original-resolution DomainNet images (Peng et al., 2019) using the Mini DomainNet image lists (Zhou et al., 2021), spanning four domains: Clipart, Painting, Real, and Sketch. We retain classes with at least 20 images in every domain, select the 50 classes with the largest total image counts across the four domains, and draw a class-proportional stratified sample of 5,000 images independently within each domain. The complete class list and class-wise image counts are provided in Appendix E.3.2.

Baselines. For a fair comparison, FL baselines use the same heterogeneous-defense setting: the defended client uses AT, while weakly defended clients use clean CE unless specified otherwise. We prefix them with **SR** to denote selective robust training. **Robust ERM** performs adversarial training on a single defended source client without federated aggregation. **SR-FedAvg** (McMahan et al., 2017) aggregates the selectively trained client models using standard sample-count weights. **SR-FedProx** (Li et al., 2020) adds proximal regularization ($\mu = 0.1$) to SR-FedAvg. **SR-FRP** (Hong et al., 2023), originally designed for target-aware robustness propagation to participating low-resource clients, is adapted to our target-excluded LODO setting by replacing its mixed clean/adversarial objective with standard AT, while retaining its dual-BN structure and $\lambda_{\mathrm{FRP}} = 0.5$. Because target-domain BN statistics are unavailable, the global model uses the defended client's noise-BN parameters.

Training Details. All methods use a ResNet-50 (He et al., 2016) initialized from Madry Lab's ℓ_∞-robust ImageNet model (Salman et al., 2020), pretrained at $\varepsilon = 1/255$. We follow DomainBed's fixed-step convention (Gulrajani & Lopez-Paz, 2021), using SGD with learning rate 0.001, batch size 32, and $E = 20$ local updates per round for 250 rounds. Adversarial training uses PGD-10 with $\varepsilon \in \{2/255, 4/255, 8/255\}$ and step size $\varepsilon/4$. FedR^{2}DG uses $T_{\mathrm{warm}} = 50$ clean warm-up rounds (1,000 local steps) before $T = 200$ main-training rounds. Weakly defended clients use

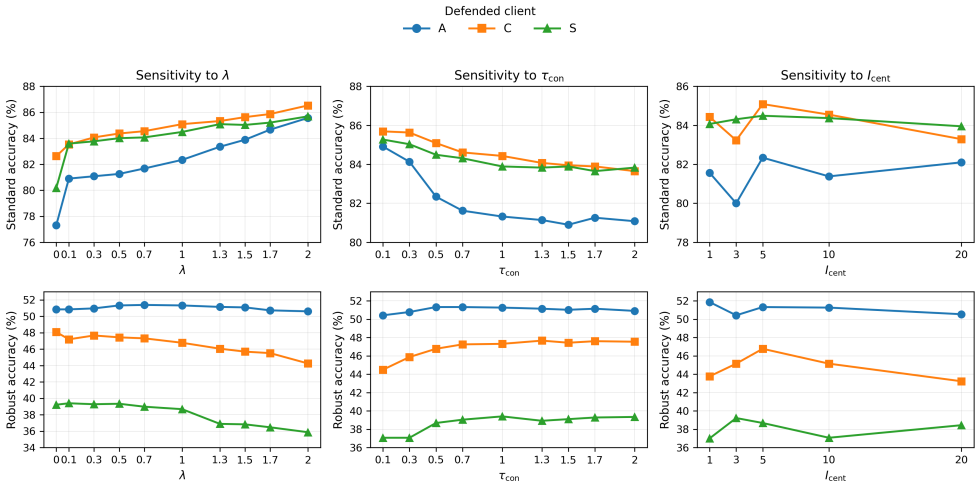


Figure 2: Standard and robust accuracy as a function of the prototype loss weight λ, temperature τ_{con}, and centroid update interval I_{cent}, across defended-client choices.

$(B_C, B_H, \zeta) = (5, 5, 3)$ per main-training round. We set $\gamma_{\text{ema}} = 0.1$, $\alpha_{\text{ema}} = 0.9$, $I_{\text{cent}} = 5$, $\lambda = 1.0$, and $\tau_{\text{con}} = 0.5$.

5.2 Performance Evaluation

Robust evaluation is conducted against AutoAttack (AA) (Croce & Hein, 2020) at the same ε values, and results are averaged over three seeds. Table 1 reports standard accuracy, robust accuracy against AA, and H-score, $H = 2A_{\text{std}}A_{\text{rob}}/(A_{\text{std}} + A_{\text{rob}})$, where A_{std} and A_{rob} are standard and robust accuracy. We compute H per case and seed, average the 12 cases within each seed, then average over seeds. Appendix D provides full per-case results, while Appendix C shows that clean fine-tuning retains little robustness at the largest attack budget. Across the 108 combinations of dataset, held-out target, defended client, and attack budget, FedR^{2}DG ranks first in 105 cases by three-seed mean robust accuracy and in 105 cases by mean H-score; the exceptions differ between the two metrics.

Standard–robust balance. FedR^{2}DG achieves the highest aggregate robust accuracy and H-score across all nine dataset/budget settings while largely preserving standard accuracy. Compared with SR-FedAvg, its standard accuracy remains similar across perturbation budgets. SR-FRP often attains higher standard accuracy, particularly at larger budgets, but provides weaker robust transfer to the held-out domain, consistent with the standard–robust trade-off (Tsipras et al., 2019; Zhang et al., 2019). Absolute robust accuracy decreases for all methods as attacks become stronger, but FedR^{2}DG retains its advantage. We observe no catastrophic overfitting (Wong et al., 2020) in any run.

5.3 Ablation Studies

All ablations are conducted on PACS with P as the target domain, while the defended client varies among A, C, and S, as shown in Figure 2.

Hyperparameter Sensitivity. We sweep the prototype weight λ over ten values in $[0, 2]$, including the no-prototype case $\lambda = 0$, and separately sweep the temperature τ_{con} over nine values in $[0.1, 2]$. Increasing λ monotonically improves standard accuracy across all three cases, while robust accuracy remains stable up to about $\lambda = 1.0$ and then declines as the prototype term becomes dominant. This confirms that the prototype loss acts as a controllable clean-feature anchor without an immediate robustness cost. For τ_{con}, larger values generally reduce standard accuracy, whereas robust accuracy stays roughly flat or improves before plateauing. Although these hyperparameters could be tuned per case, all main results use the single fixed pair $\lambda = 1.0$ and $\tau_{\text{con}} = 0.5$.

Table 2: Computational cost per client in TMACs. Relative cost denotes FedR^{2}DG/SR-FedAvg. Main-round training excludes centroid computation; full-run costs include warm-up, centroid initialization, and 40 refreshes ($I_{\text{cent}} = 5$). Centroid and full-run costs use each dataset's mean domain size.

Scope	Cost item	FedR^{2}DG	SR-FedAvg	Ratio
Main round	Defended	94.68	86.79	1.09×
	Weakly defended	7.89	7.89	1.00×
PACS	Centroid/call	10.52	–	–
	Defended total	19,761.84	21,697.52	0.91×
	Weakly def. total	2,403.82	1,972.50	1.22×
OfficeHome	Centroid/call	16.04	–	–
	Defended total	19,988.28	21,697.52	0.92×
	Weakly def. total	2,630.27	1,972.50	1.33×
mini-DomainNet50	Centroid/call	20.78	–	–
	Defended total	20,182.37	21,697.52	0.93×
	Weakly def. total	2,824.36	1,972.50	1.43×

Centroid Update Interval. We compare $I_{\text{cent}} \in \{1, 3, 5, 10, 20\}$. Setting $I_{\text{cent}} = 5$ is comparable to or better than $I_{\text{cent}} = 1$ in most cases, while larger intervals show no consistent additional gains. Combined with the TMAC-based cost analysis in Section 5.4, $I_{\text{cent}} = 5$ provides a practical balance between performance and the compute savings from less frequent centroid updates.

5.4 Computational Cost Analysis

Table 2 compares per-client costs with SR-FedAvg in tera multiply–accumulate operations (TMACs), focusing on weakly defended clients under heterogeneous budgets.

Excluding centroid computation, a weakly defended client's per-round cost is identical across both methods: both perform 20 local updates, and replay obtains parameter and input gradients from a single backward pass (Section 4.2). Both methods use PGD-10 for defended clients; FedR^{2}DG's additional prototype loss increases their main-round cost, while its 50 clean warm-up rounds lower their full-run cost. Centroids are recomputed without gradients every $I_{\text{cent}} = 5$ rounds (40 refreshes over 200 main rounds). Averaged across the three datasets using their mean domain sizes, a weakly defended client's centroid-update round costs 23.67 TMACs, about 25% of a defended-client round. Including warm-up, centroid initialization, and refreshes, the weakly defended client's total cost is 1.22–1.43× that of SR-FedAvg, with the overhead attributable to centroid computation. Less frequent centroid updates reduce this overhead (Section 5.3).

6 Conclusions

We introduced selective federated robust domain generalization, where clients differ in both domain and defense budget and the target domain remains unseen during training. Our held-out robust-risk analysis shows that robust transfer depends on source-domain robust fit, federation-induced drift, class-conditional source coverage, label-prior mismatch, and a conditional residual. We bound the residual by source conditional hull approximation and mixture-weight mismatch; it vanishes when the target class-conditionals match the aggregation-induced mixture. FedR^{2}DG follows this decomposition by combining selective local defense with drift control, class-conditional alignment, and aggregation reweighting. Experiments show that FedR^{2}DG improves robust transfer and the standard–robust balance with modest additional cost for weakly defended clients, suggesting that full adversarial training is not required at every client.

Limitations Despite these gains, the conditional residual is not directly observable or optimizable without target-domain data. The residual bound controls it through source conditional hull approximation and mixture-weight mismatch, but may remain loose when the target lies far from this hull. Label-prior mismatch is addressed only indirectly through aggregation weights, and larger federations or different defense ratios remain future work.

Reproducibility Statement

The assumptions and proof are provided in Appendix A, and Appendix B specifies the training procedure. Appendix E documents the experimental environment, seeds, hyperparameters, and dataset construction; Appendix D reports results for each target and defended-client choice.

AI Use Statement

Generative AI tools (OpenAI Codex) assisted with drafting and polishing parts of the manuscript, manuscript organization, related-work search, notation standardization, mathematical consistency checks, and LaTeX layout and compilation. These tools also assisted in checking the correspondence between reported results and their written interpretation. The authors reviewed all AI-assisted outputs and are responsible for the final content of this submission.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. URL https://www.jmlr.org/papers/v3/bartlett02a.html.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. URL https://jmlr.org/papers/v22/17-679.html.

Chen Chen, Yuchen Liu, Xingjun Ma, and Lingjuan Lyu. Calfat: Calibrated federated adversarial training with label skewness. *Advances in Neural Information Processing Systems*, 35:3569–3581, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/171c3678c36e39fc0074f3e7332a9a66-Abstract-Conference.html.

Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12077–12086, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_Fair_Federated_Learning_under_Domain_Skew_with_Local_Consistency_and_CVPR_2024_paper.html.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 2017.

Corinna Cortes, Mehryar Mohri, Ananda Theertha Suresh, and Ningshan Zhang. A discriminative technique for multiple-source adaptation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2132–2143. PMLR, 2021. URL https://proceedings.mlr.press/v139/cortes21b.html.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2705–2718, 2022.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

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.

Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: Improving federated learning on heterogeneous data via local learning bias reduction. In *International conference on machine learning*, pp. 12034–12054. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated robustness propagation: Sharing adversarial robustness in heterogeneous federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):7893–7901, 2023. doi: 10.1609/aaai.v37i7.25955. URL https://ojs.aaai.org/index.php/AAAI/article/view/25955.

Kasra Jalaldoust, Alexis Bellot, and Elias Bareinboim. Partial transportability for domain generalization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4376.

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. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826. PMLR, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. URL https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

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.

Xiaoxiao Li, Zhao Song, and Jiaming Yang. Federated adversarial learning: A framework with convergence analysis. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19932–19959. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23z.html.

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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7313–7324. PMLR, 2021. URL https://proceedings.mlr.press/v139/mahajan21b.html.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

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.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. doi: 10.1214/12-EJS722. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-6/issue-none/On-the-empirical-estimation-of-integral-probability-metrics/10.1214/12-EJS722.full.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

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 (CVPR)*, pp. 5018–5027, 2017.

Xi Wang, Liang Bai, Xian Yang, and Jiye Liang. Expanding domain generalization theory: Error bound beyond convex combination assumption. *Machine Learning*, 115, 2026. doi: 10.1007/s10994-026-07003-3. Article 98.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

Dongyoon Yang, Insung Kong, and Yongdai Kim. Improving adversarial robustness by putting more regularizations on less robust samples. In *International Conference on Machine Learning*, pp. 39331–39348. PMLR, 2023.

Dongyoon Yang, Jihu Lee, and Yongdai Kim. TAROT: Towards essentially domain-invariant robustness with theoretical justification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25780–25789, 2025.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7085–7094. PMLR, 2019. URL https://proceedings.mlr.press/v97/yin19b.html.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

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 (CVPR)*, pp. 3954–3963, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Federated_Domain_Generalization_With_Generalization_Adjustment_CVPR_2023_paper.html.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.

Xin Zou and Weiwei Liu. On the adversarial robustness of out-of-distribution generalization models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=IiwTFcGGTq.

A Theoretical Analysis

We give the notation, assumptions, finite-sample bounds, and proofs.

A.1 Notation Reference

Table 3: Notation reference. Hats on losses denote empirical counterparts. Source laws, sample sizes, attack radius, temperature, and function classes are fixed.

Symbol	Meaning
$K, K_s, C;\ i, j, c$	Domain, source-client, and class counts; client, target, and class indices.
$S_j, \mathcal{D}_j, \mathcal{N}_j$	Source, defended, and weakly defended client sets for target j.
$\mathcal{P}_i, \widehat{D}_i; n_i, n_{i,c}$	Client distribution, observed sample, total count, and class count.
$\nu_i, \mathcal{P}_\nu$	Model aggregation weight and weighted source distribution.
$\pi_i, \pi_\nu; \widehat{\pi}_i, \widehat{\pi}_\nu$	Population label priors and empirical estimates.
$\nu_{i,c}, p_{i,c}$	Population source weight conditional on class c; empirical class-count weight for centroids.
$\boldsymbol{\theta}_i^{(t)}, \boldsymbol{b}_i^{(t)}, \boldsymbol{\omega}_i^{(t)}$	Trainable parameters, prediction buffers, and complete state $\boldsymbol{\omega}_i^{(t)} = (\boldsymbol{\theta}_i^{(t)}, \boldsymbol{b}_i^{(t)})$.
$\bar{\boldsymbol{\theta}}^{(t)}, \bar{\boldsymbol{b}}^{(t)}, \bar{\boldsymbol{\omega}}^{(t)}$	Server averages of the corresponding client states.
$h_i^{(t)}, \bar{h}^{(t)}$	Returned local predictor and aggregated predictor; $h_{\boldsymbol{\theta}}$ suppresses buffers in training notation.
$\boldsymbol{\phi}, \mathbf{Z}_{\boldsymbol{\phi}}; \boldsymbol{\phi}_{\boldsymbol{\theta}}, \mathbf{Z}_{\boldsymbol{\theta}}$	Analyzed global feature map and normalized feature; evolving local counterparts.
$v_{i,c}, v_c, \mathbf{v}$	Local class centroid, global class anchor, and collection of global anchors.
$\varepsilon, \boldsymbol{\delta}; \beta$	Attack radius and perturbation; failure-probability parameter.
$g_h(\boldsymbol{x}, y), \mathcal{R}_i^{\varepsilon}(h)$	Pointwise robust loss and population robust risk.
B, M_{proto}	Robust- and prototype-loss upper bounds; $M_{\mathrm{proto}} = \log(1 + (C-1)e^{2/\tau_{\mathrm{con}}})$.
$\tau_{\mathrm{con}}, \lambda$	Prototype temperature and weight; $\tau = \tau_{\mathrm{con}}$ in calibration.
$T_{\mathrm{warm}}, T; t, k$	Warm-up rounds, main rounds, round index, and local-step index.
$E_i, E; \eta_i, g_i^{(t,k)}, G_i$	Local steps (common budget E), learning rate, effective optimizer direction, and its norm bound.
B_C, B_H, ζ	Clean steps, replay mini-batches, and replays per batch.
$\gamma_i, \gamma_{\mathrm{ema}}$	Client-specific tether coefficient in the proof and common experimental value.
$\alpha_{\mathrm{ema}}, I_{\mathrm{cent}}$	Centroid EMA coefficient and refresh interval in rounds.
Bound quantities and their dependence	
$L_{\mathrm{src}}(\mathbf{h}; \nu)$	Population robust fit of local predictors $\mathbf{h} = (h_i)_{i \in S_j}$ with source weights ν.
$L_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu)$	Population prototype loss for a shared representation, anchors, and source weights.
$L_{\mathrm{proto},c}(\boldsymbol{\phi}, \mathbf{v}; \nu)$	Class-c prototype loss, averaged with conditional weights $\nu_{i,c}$.
$r_Y(\nu), \widehat{r}_Y(\nu)$	Population and empirical source label-prior radii; independent of learned features.
$\mathfrak{R}_{n,\mathcal{P}}(\mathcal{G})$	Expected Rademacher complexity of function class $\mathcal{G}$ for n samples from $\mathcal{P}$.
$\mathcal{G}_{\mathrm{rob}}, \mathcal{G}_{\mathrm{proto}}$	Fixed classes of robust and prototype losses over all allowed models and anchors.
$\mathcal{E}_{\mathrm{src}}(\beta; \nu)$	Source robust-loss error bound: weighted Rademacher complexity plus concentration.
$\mathcal{E}_{\mathrm{proto}}(\beta; \nu)$	Prototype-loss error bound: weighted Rademacher complexity plus concentration.
$\mathcal{E}_{\mathrm{prior}}(\beta)$	Label-prior error bound from concentration; depends on C, sample sizes, and confidence.
$\mathcal{A}_\beta$	Event where all three error bounds hold uniformly; probability at least $1 - 3\beta$.
$\Delta_i^{(T)}, \boldsymbol{\Delta}^{(T)}$	Complete-state drift from the broadcast state; collection of client drifts.
$U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)}; \nu)$	Weighted robust-risk penalty for client drift.
$L_{\mathrm{stab}}, M_\varepsilon$	Fixed risk-to-function and state-to-function stability constants.
$q_{r,c}^h(z; \boldsymbol{\phi})$	Conditional robust-loss profile for law $\mathcal{P}_r$, class c, predictor h, and representation $\boldsymbol{\phi}$.
$\rho_j(h; \boldsymbol{\phi}, \nu)$	Conditional target–source residual for the predictor, representation, and mixture.
$\mathcal{F}, L_{\mathcal{F}}$	Fixed discrepancy class and its critic Lipschitz constant.
$D_{\mathrm{src}}(\boldsymbol{\phi}; \nu)$	Source class-conditional IPM dispersion in the learned representation.
$C_{\mathrm{cc}}(\boldsymbol{\phi}; \nu)$	Compatibility of source conditional profiles with the fixed discrepancy class.
$\kappa_{\mathrm{cc}}^j(\boldsymbol{\phi}; \nu)$	Target feature-coverage factor relative to source dispersion.
$\kappa_\pi^j(\nu)$	Target prior-coverage factor; independent of representation and anchors.
$\xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu)$	Source imbalance of class-c dispersion around its anchor; $\xi_{\max}(\boldsymbol{\phi}, \mathbf{v}; \nu)$ is the maximum over classes.
$\Lambda_{\mathrm{cov}}^j(\boldsymbol{\phi}, \mathbf{v}; \nu)$	Combined coefficient for target coverage, profile compatibility, and centroid imbalance.
$\Psi_{\tau,C}(u), b_{\tau,C}$	Prototype-loss calibration map and its domain's lower endpoint.

A.2 Regularity Conditions and Diagnostics

Common convention and auxiliary notation. The proof allows E_i and γ_i to vary by client; the experiments use $E_i = E$ and $\gamma_i = \gamma_{\mathrm{ema}}$. Throughout the appendix, representation-dependent quantities are evaluated for a fixed analyzed round. The source laws and sample sizes, ε, τ_{con}, and model/loss classes are fixed. In the main theorem, ϕ is the feature extractor of the round-T aggregated model $\bar{h}^{(T)}$. Definitions, assumptions, and proof steps retain the same functional arguments as the main bound. Set $\tau = \tau_{\mathrm{con}}$ and let

$$\mathbf{Z} := \mathbf{Z}_{\phi}(\mathbf{X}) := \frac{\phi(\mathbf{X})}{\|\phi(\mathbf{X})\|_2} \in \mathbb{S}^{d-1}$$

denote the normalized representation, and let $\mathbf{v} = (v_1, \ldots, v_C)$ collect the normalized global class centroids in the same representation space. Define the source-supported and zero-source-mass class sets by

$$\mathcal{C}_\nu := \{c \in [C] : \pi_\nu(c) > 0\},\ \mathcal{C}_\nu^0 := [C] \setminus \mathcal{C}_\nu.$$

All source-mixture class-conditional quantities are evaluated only for $c \in \mathcal{C}_\nu$. Classes in $\mathcal{C}_\nu^0$ are omitted from source-mixture class-wise sums and are controlled through the label-prior term. If a class-conditional term is multiplied by a zero class probability, the corresponding term is interpreted as zero. Post-hoc checks involving the held-out domain j are used only for dataset characterization, not for training or model selection.

Let d_ε denote the robust function-space pseudometric used to compare two predictors under perturbation radius ε.

For $c \in \mathcal{C}_\nu$, define

$$S_j(c) := \{i \in S_j : \pi_i(c) > 0\},\ \nu_{i,c} := \frac{\nu_i \pi_i(c)}{\pi_\nu(c)}.$$

The class-conditional source barycenter in representation space is

$$\mathcal{P}_{\nu,\mathbf{Z}|Y=c} := \sum_{i \in S_j(c)} \nu_{i,c} \mathcal{P}_{i,\mathbf{Z}|Y=c}.$$

For $r \in \{j, \nu\}$, define the class-conditional robust profile

$$q^h_{r,c}(z;\phi) := \mathbb{E}_{(\mathbf{X},Y)\sim\mathcal{P}_r}\big[g_h(\mathbf{X},Y) \mid \mathbf{Z} = z, Y = c\big],$$

where the conditioning on $\mathbf{Z} = z$ is understood in the regular-conditional sense. Input and feature spaces are standard Borel spaces, and g_h and $\mathbf{Z}_\phi$ are measurable. We fix one measurable version of each $q^h_{r,c}(z;\phi)$ throughout, with a fixed bounded extension on null sets of its conditioning law. The source version must also satisfy Assumption 2. The residual can depend on this extension when the target feature law charges a source-null set, and is not a source-identifiable quantity. No absolute continuity of the target law with respect to the source law is assumed. For Lemma 1 and its corollaries, the chosen source profiles must additionally take values in $[0, B]$ on these null sets. Bounded robust loss ensures this range only almost surely under the source conditioning law. This extra condition is not imposed on Theorem 1; its label-prior step uses only source expectations.

Define the label-prior source radius

$$r_Y(\nu) := \max_{i \in S_j} \|\pi_i - \pi_\nu\|_1.$$

For two distributions P and Q on the representation space, define

$$\mathrm{IPM}_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\right|.$$

The class-conditional source dispersion is

$$D_{\mathrm{src}}(\phi;\nu) := \sum_{c \in \mathcal{C}_\nu} \pi_\nu(c) \max_{i \in S_j(c)} \mathrm{IPM}_{\mathcal{F}}\big(\mathcal{P}_{i,\mathbf{Z}|Y=c}, \mathcal{P}_{\nu,\mathbf{Z}|Y=c}\big).$$

Here $D_{\mathrm{src}}(\phi;\nu)$ measures geometric dispersion via an IPM; it is distinct from the Rademacher complexity defined in Appendix A.3. The critic class $\mathcal{F}$ is fixed.

The class-conditional residual is

$$\rho_j(h;\phi,\nu) := \left| \sum_{c \in \mathcal{C}_\nu} \pi_j(c) \mathbb{E}_{\mathbf{Z}\sim\mathcal{P}_{j,\mathbf{Z}|Y=c}}\left[q^h_{j,c}(\mathbf{Z};\phi) - q^h_{\nu,c}(\mathbf{Z};\phi)\right] \right|.$$

Loss convention. We work with a clipped robust loss, or an intrinsically bounded surrogate, so there exists $B > 0$ such that, for every $h \in \mathcal{H}$ and every input-label pair $(\boldsymbol{x}, y)$,

$$0 \le g_h(\boldsymbol{x}, y) \le B.$$

This is an enforceable modeling convention rather than a source-coverage condition. The constant B scales the label-prior mismatch term and the concentration term for robust loss. Label-marginal extrapolation cannot be controlled by representation alignment alone.

Assumption 1. There exist $\kappa_{\mathrm{cc}}^{j}(\boldsymbol{\phi}; \nu) \ge 1$ and $\kappa_{\pi}^{j}(\nu) \ge 1$ such that

$$\sum_{c \in \mathcal{C}_\nu} \pi_j(c)\, \mathrm{IPM}_{\mathcal{F}}\big(\mathcal{P}_{j,\mathbf{Z}|Y=c}, \mathcal{P}_{\nu,\mathbf{Z}|Y=c}\big) \le \kappa_{\mathrm{cc}}^{j}(\boldsymbol{\phi}; \nu) D_{\mathrm{src}}(\boldsymbol{\phi}; \nu), \tag{13}$$

and

$$\|\pi_j - \pi_\nu\|_1 \le \kappa_{\pi}^{j}(\nu) r_Y(\nu). \tag{14}$$

Validation and references. This condition states that the held-out domain is not a severe extrapolation beyond the source federation in either class-conditional representation geometry or label-prior geometry. It is a modeling assumption of this paper, motivated by source-mixture geometry in multiple-source adaptation (Cortes et al., 2021) and risk extrapolation in DG (Krueger et al., 2021); these works do not establish the stated coverage inequalities for arbitrary held-out domains. Label-prior mismatch warrants separate treatment: invariant representations alone need not ensure low source and target errors when label marginals differ (Zhao et al., 2019), and label skew also affects federated adversarial training (Chen et al., 2022). In leave-one-domain-out evaluation, the empirical ratios corresponding to Equations (13) and (14) can be reported post hoc to separate representation shift from class-prior mismatch. These diagnostics are used only for dataset characterization, not for training or model selection.

Assumption 2. There exist finite constants $C_{\mathrm{cc}}(\boldsymbol{\phi}; \nu) > 0$ and $L_{\mathcal{F}} > 0$, where $C_{\mathrm{cc}}(\boldsymbol{\phi}; \nu)$ is a class-conditional compatibility constant, such that the following hold. First, for every $h \in \mathcal{H}$ and every class $c \in \mathcal{C}_\nu$, the class-conditional robust profile $z \mapsto q_{\nu,c}^{h}(z; \boldsymbol{\phi})$ is dominated by the IPM class $\mathcal{F}$:

$$\frac{q_{\nu,c}^{h}(\,\cdot\,; \boldsymbol{\phi})}{C_{\mathrm{cc}}(\boldsymbol{\phi}; \nu)} \in \mathcal{F}.$$

Consequently, for any distributions P, Q on the representation space,

$$\big|\mathbb{E}_{\mathbf{Z}\sim P} q_{\nu,c}^{h}(\mathbf{Z}; \boldsymbol{\phi}) - \mathbb{E}_{\mathbf{Z}\sim Q} q_{\nu,c}^{h}(\mathbf{Z}; \boldsymbol{\phi})\big| \le C_{\mathrm{cc}}(\boldsymbol{\phi}; \nu)\, \mathrm{IPM}_{\mathcal{F}}(P, Q).$$

Second, every $f \in \mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz:

$$|f(z) - f(z')| \le L_{\mathcal{F}} \|z - z'\|_2, \qquad \forall z, z' \in \mathbb{R}^d.$$

Validation and references. This is a modeling condition on the discrepancy class, not an additional held-out-domain coverage requirement. It is standard in IPM-based generalization arguments: choosing $\mathcal{F}$ as a bounded Lipschitz critic gives a Wasserstein-style discrepancy, while choosing an RKHS unit ball gives an MMD-style discrepancy (Sriperumbudur et al., 2012). In implementation, the Lipschitz part can be enforced or approximately controlled by spectral normalization, gradient penalties, or bounded-norm critic architectures. One can also check that the coverage term is qualitatively stable across reasonable choices of $\mathcal{F}$.

Centroid imbalance factor. For each class $c \in \mathcal{C}_\nu$, we expose the source-only heterogeneity around the global centroid through a finite factor $\xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu) \ge 1$ satisfying

$$\max_{i \in S_j(c)} \mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}} \|\mathbf{Z} - v_c\|_2^2 \le \xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu) \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}} \|\mathbf{Z} - v_c\|_2^2.$$

Rather than treating this as a separate unseen-domain condition, $\xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu)$ is a measurable source imbalance constant: large values mean that at least one source domain has much larger class-conditional dispersion than the source mixture. Positive aggregation weights ensure a finite choice, for example $\xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu) \le 1/\min_{i \in S_j(c)} \nu_{i,c}$ can be achieved. This is natural for prototype- or centroid-based class alignment, where compact class-wise representations are encouraged around shared prototypes (Snell et al., 2017; Khosla et al., 2020). Since $\xi_c(\boldsymbol{\phi}, \mathbf{v}; \nu)$ depends only on source representations, large values can be inspected through class-wise centroid dispersion and source support counts.

Assumption 3. There exist $L_{\mathrm{stab}} > 0$, $M_\varepsilon > 0$, and client-wise update bounds $G_i > 0$ such that, for every client i and every $h, h' \in \mathcal{H}$,

$$|\mathcal{R}_i^\varepsilon(h) - \mathcal{R}_i^\varepsilon(h')| \le L_{\mathrm{stab}}\, d_\varepsilon(h, h'),$$

for all complete prediction states $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{b})$ and $\boldsymbol{\omega}' = (\boldsymbol{\theta}', \boldsymbol{b}')$,

$$d_\varepsilon(h_{\boldsymbol{\omega}}, h_{\boldsymbol{\omega}'}) \le M_\varepsilon \|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_2,$$

and each effective local optimizer direction satisfies

$$\|g_i^{(T,k)}\|_2 \le G_i.$$

Here E_i is the number of local optimizer steps taken by client i in round T, and $\gamma_i \in [0, 1]$ is its EMA tethering coefficient.

Validation and references. This condition packages the regularity needed to convert local federated drift into a robust-risk gap. The first two inequalities hold when the robust loss and the network have controlled Lipschitz behavior under the pseudometric d_ε; spectral normalization and Parseval-style networks are representative mechanisms for enforcing such stability (Miyato et al., 2018; Cisse et al., 2017). The update bound is operational and can be enforced by clipping the effective optimizer direction (after momentum and weight decay),

$$g_i^{(T,k)} \leftarrow G_i \, \frac{g_i^{(T,k)}}{\max\{\|g_i^{(T,k)}\|_2, G_i\}}.$$

Bounded-gradient or bounded-update conditions are common in federated optimization analyses under heterogeneity, including FedProx- and SCAFFOLD-style analyses (Li et al., 2020; Karimireddy et al., 2020). Empirically, this condition is tied to measurable training quantities such as the clipping threshold, the fraction of clipped updates, and the scale of local parameter drift.

A.3 Empirical Quantities and Rademacher Generalization Bounds

Let $\widehat{D}_i = \{(\boldsymbol{x}_{i,m}, y_{i,m})\}_{m=1}^{n_i}$ be the observed source sample on client i. The empirical robust risk is

$$\widehat{\mathcal{R}}_i^\varepsilon(h) := \frac{1}{n_i} \sum_{m=1}^{n_i} g_h(\boldsymbol{x}_{i,m}, y_{i,m}).$$

The population and empirical source local-risk terms are

$$L_{\mathrm{src}}(\mathbf{h}^{(T)}; \nu) := \sum_{i \in S_j} \nu_i \mathcal{R}_i^\varepsilon(h_i^{(T)}), \qquad \widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)}; \nu) := \sum_{i \in S_j} \nu_i \widehat{\mathcal{R}}_i^\varepsilon(h_i^{(T)}).$$

For the global class-centroid contrastive loss, define

$$L_{\mathrm{proto},c}(\boldsymbol{\phi}, \mathbf{v}; \nu) := \sum_{i \in S_j(c)} \nu_{i,c} \mathbb{E}_{\mathbf{X} \sim \mathcal{P}_{i,\mathbf{X}|Y=c}} \big[\ell_{\mathrm{proto}}(\mathbf{X}, c; \boldsymbol{\phi}, \mathbf{v})\big],$$

and

$$L_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu) := \sum_{c \in \mathcal{C}_\nu} \pi_\nu(c) L_{\mathrm{proto},c}(\boldsymbol{\phi}, \mathbf{v}; \nu).$$

Let $\widehat{\pi}_i(c) := n_{i,c}/n_i$, $\widehat{\pi}_\nu(c) := \sum_{i \in S_j} \nu_i \widehat{\pi}_i(c)$, and $\widehat{\mathcal{C}}_\nu := \{c : \widehat{\pi}_\nu(c) > 0\}$. For $c \in \widehat{\mathcal{C}}_\nu$, set

$$\widehat{\nu}_{i,c} := \frac{\nu_i \widehat{\pi}_i(c)}{\widehat{\pi}_\nu(c)}.$$

The empirical centroid contrastive loss is

$$\widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu) := \sum_{c \in \widehat{\mathcal{C}}_\nu} \widehat{\pi}_\nu(c) \sum_{i : \widehat{\pi}_i(c) > 0} \widehat{\nu}_{i,c} \frac{1}{n_{i,c}} \sum_{m : y_{i,m} = c} \ell_{\mathrm{proto}}(\boldsymbol{x}_{i,m}, c; \boldsymbol{\phi}, \mathbf{v}).$$

Since $\widehat{\pi}_\nu(c)\widehat{\nu}_{i,c}/n_{i,c} = \nu_i/n_i$ for observed client classes, this equals the simple weighted sample average in Equation (1).

The empirical label-prior radius is

$$\widehat{r}_Y(\nu) := \max_{i \in S_j} \|\widehat{\pi}_i - \widehat{\pi}_\nu\|_1.$$

For a fixed function class $\mathcal{G}$, sampling law $\mathcal{P}$, and i.i.d. sample $\widehat{D} = (u_m)_{m=1}^n \sim \mathcal{P}^n$, define the expected Rademacher complexity

$$\mathfrak{R}_{n,\mathcal{P}}(\mathcal{G}) := \mathbb{E}_{\widehat{D}\sim\mathcal{P}^n,\sigma}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{m=1}^n \sigma_m g(u_m)\right],$$

where the signs σ_m are independent, uniform on $\{-1, 1\}$, and independent of the sample. Thus $\mathfrak{R}_{n,\mathcal{P}}(\mathcal{G})$ depends on the sample size, sampling law, and entire function class; it is not the loss of a particular fitted model.

For a failure probability $\beta \in (0, 1/3)$, the corrections $\mathcal{E}_{\mathrm{src}}(\beta;\nu)$, $\mathcal{E}_{\mathrm{proto}}(\beta;\nu)$, and $\mathcal{E}_{\mathrm{prior}}(\beta)$ defined below satisfy, with probability at least $1 - 3\beta$ over the source samples,

$$L_{\mathrm{src}}(\mathbf{h}^{(T)};\nu) \le \widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)};\nu) + \mathcal{E}_{\mathrm{src}}(\beta;\nu), \tag{15}$$

$$L_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v};\nu) \le \widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v};\nu) + \mathcal{E}_{\mathrm{proto}}(\beta;\nu), \tag{16}$$

and

$$r_Y(\nu) \le \widehat{r}_Y(\nu) + \mathcal{E}_{\mathrm{prior}}(\beta). \tag{17}$$

Let $\mathcal{A}_\beta$ denote the event on which Equations (15) to (17) hold uniformly over the allowed predictors, feature maps, and anchors. The first two corrections combine Rademacher complexity with concentration; the third controls estimation of label proportions. This distinction allows models and centroids to be selected using the same source data while keeping the classes fixed before sampling.

Rademacher complexity and concentration terms. Assume each $\widehat{D}_i$ consists of n_i i.i.d. samples from $\mathcal{P}_i$, with sample sizes and aggregation weights fixed conditional on the sampling design. Set $K_s = |S_j|$, and fix $\tau_{\mathrm{con}} > 0$. Let $\mathcal{H}$ and Φ be the prescribed predictor and feature-map classes, and define

$$\mathcal{G}_{\mathrm{rob}} = \{g_h : h \in \mathcal{H}\}, \qquad \mathcal{G}_{\mathrm{proto}} = \{\ell_{\mathrm{proto}}(\,\cdot\,;\boldsymbol{\phi}, \mathbf{v}) : \boldsymbol{\phi} \in \Phi,\ \mathbf{v} \in (\mathbb{S}^{d-1})^C\}.$$

Their ranges are $[0, B]$ and $[0, M_{\mathrm{proto}}]$, respectively, where

$$M_{\mathrm{proto}} = \log\bigl(1 + (C-1)e^{2/\tau_{\mathrm{con}}}\bigr).$$

The complexity must be evaluated on the adversarially maximized loss class $\mathcal{G}_{\mathrm{rob}}$, as in robust Rademacher analyses (Yin et al., 2019). Robust learning can require more samples than standard learning (Schmidt et al., 2018); replacing this class by its clean-loss counterpart would require a separate justification. Symmetrization, bounded-difference concentration, and a union bound over clients give the valid choices (Bartlett & Mendelson, 2002; Talagrand, 2014)

$$\mathcal{E}_{\mathrm{src}}(\beta;\nu) = \sum_{i\in S_j} \nu_i \left[2\mathfrak{R}_{n_i,\mathcal{P}_i}(\mathcal{G}_{\mathrm{rob}}) + B\sqrt{\frac{\log(K_s/\beta)}{2n_i}}\right],$$

$$\mathcal{E}_{\mathrm{proto}}(\beta;\nu) = \sum_{i\in S_j} \nu_i \left[2\mathfrak{R}_{n_i,\mathcal{P}_i}(\mathcal{G}_{\mathrm{proto}}) + M_{\mathrm{proto}}\sqrt{\frac{\log(K_s/\beta)}{2n_i}}\right].$$

Each bracket contains twice the expected Rademacher complexity plus a confidence-dependent concentration term. The robust-loss class depends on ε and $\mathcal{H}$; the prototype class depends on $\tau_{\mathrm{con}}, C, \Phi$ and the anchor space. These background choices, $\mathcal{P}_i$, and n_i are fixed; the displayed corrections expose β and ν. Each correction holds with probability at least $1 - \beta$, uniformly over the allowed predictors, representations, and centroids.

Label-prior concentration. For the label prior, a direct finite-class concentration bound suffices; $\mathcal{E}_{\text{prior}}(\beta)$ does not denote a Rademacher complexity. The identity $\|a\|_1 = \max_{s\in\{-1,1\}^C} s^\top a$ and Hoeffding's inequality (Hoeffding, 1963) imply

$$\|\widehat{\pi}_i - \pi_i\|_1 \le e_i(\beta), \qquad e_i(\beta) = \min\left\{2, \sqrt{2[C\log 2 + \log(K_s/\beta)]/n_i}\right\}$$

simultaneously over clients with probability at least $1-\beta$. The triangle inequality then gives

$$r_Y(\nu) \le \widehat{r}_Y(\nu) + \mathcal{E}_{\text{prior}}(\beta), \qquad \mathcal{E}_{\text{prior}}(\beta) = 2\max_{i\in S_j} e_i(\beta).$$

A final union bound yields $\Pr(\mathcal{A}_\beta) \ge 1-3\beta$. These expressions provide valid, potentially loose bounds; the experiments do not claim to estimate the Rademacher complexities or certify target risk.

Lemma 2. *Suppose every $f \in \mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz. Then, for every class $c \in \mathcal{C}_\nu$,*

$$\max_{i\in S_j(c)} \text{IPM}_{\mathcal{F}}\big(\mathcal{P}_{i,\mathbf{Z}|Y=c}, \mathcal{P}_{\nu,\mathbf{Z}|Y=c}\big) \le 2L_{\mathcal{F}} \max_{i\in S_j(c)} \mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}}\|\mathbf{Z} - v_c\|_2.$$

Remark 3. The lemma bounds source class-conditional IPM discrepancy by distances to the shared class centroid.

Proof. Fix c and $i \in S_j(c)$. For any $f \in \mathcal{F}$, adding and subtracting $f(v_c)$ gives

$$\left|\mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}} f(\mathbf{Z}) - \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}} f(\mathbf{Z})\right| \le \mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}}|f(\mathbf{Z}) - f(v_c)| + \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}}|f(\mathbf{Z}) - f(v_c)|.$$

By Lipschitzness, this is at most

$$L_{\mathcal{F}}\left(\mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}}\|\mathbf{Z} - v_c\|_2 + \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}}\|\mathbf{Z} - v_c\|_2\right).$$

Since $\mathcal{P}_{\nu,\mathbf{Z}|Y=c}$ is the barycenter,

$$\mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}}\|\mathbf{Z} - v_c\|_2 \le \max_{k\in S_j(c)} \mathbb{E}_{\mathcal{P}_{k,\mathbf{Z}|Y=c}}\|\mathbf{Z} - v_c\|_2.$$

Taking the supremum over f and then the maximum over i proves the result. □

A.4 Proof of Main Theorem

Theorem 1. *Fix j, T and $\beta \in (0, 1/3)$. Under the bounded-loss, coverage, discrepancy, and stability conditions in Appendix A.2, on $\mathcal{A}_\beta$,*

$$\begin{aligned}\mathcal{R}_j^{\varepsilon}(\bar{h}^{(T)}) \le{}& \widehat{L}_{\text{src}}(\mathbf{h}^{(T)};\nu) + \mathcal{E}_{\text{src}}(\beta;\nu) + U_{\text{drift}}(\boldsymbol{\Delta}^{(T)};\nu) \\ &+ \Lambda^j_{\text{cov}}(\boldsymbol{\phi},\mathbf{v};\nu)\sqrt{\Psi_{\tau,C}\big(\widehat{L}_{\text{proto}}(\boldsymbol{\phi},\mathbf{v};\nu) + \mathcal{E}_{\text{proto}}(\beta;\nu)\big)} \\ &+ B\kappa^j_\pi(\nu)\big(\widehat{r}_Y(\nu) + \mathcal{E}_{\text{prior}}(\beta)\big) + \rho_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu).\end{aligned} \tag{3}$$

Remark 4. See Remark 1 for the interpretation and scope of the bound.

Proof. Let $\boldsymbol{\theta}_i^{(T)} := \boldsymbol{\theta}_i^{(T,E_i)}$ and assume $\boldsymbol{\theta}_i^{(T,0)} = \bar{\boldsymbol{\theta}}^{(T-1)}$. Define

$$\Delta_i^{(T)} := \|\boldsymbol{\omega}_i^{(T)} - \bar{\boldsymbol{\omega}}^{(T-1)}\|_2.$$

The returned model includes both trainable parameters and prediction buffers. Write $\boldsymbol{\omega}_i^{(T)} = (\boldsymbol{\theta}_i^{(T)}, \boldsymbol{b}_i^{(T)})$ and $\bar{\boldsymbol{\omega}}^{(T)} = \sum_i \nu_i \boldsymbol{\omega}_i^{(T)}$.

Step 1: risk decomposition. The contribution of classes with zero source mass is

$$M_{j,0}(\bar{h}^{(T)};\nu) := \sum_{c\in\mathcal{C}_\nu^0} \pi_j(c)\mathbb{E}_{\mathcal{P}_j}[g_{\bar{h}^{(T)}}(\mathbf{X},Y) \mid Y=c] \le B\sum_{c\in\mathcal{C}_\nu^0}\pi_j(c).$$

For the remaining classes, the law of total expectation gives

$$\mathcal{R}_j^{\varepsilon}(\bar{h}^{(T)}) = \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\mathbb{E}_{\mathcal{P}_{j,\mathbf{Z}|Y=c}}[q_{j,c}^{\bar{h}^{(T)}}(\mathbf{Z};\boldsymbol{\phi})] + M_{j,0}(\bar{h}^{(T)};\nu).$$

Adding and subtracting the source conditional profile and its source expectation yields the exact signed decomposition

$$\mathcal{R}_j^{\varepsilon}(\bar{h}^{(T)}) = \mathcal{R}_\nu^{\varepsilon}(\bar{h}^{(T)}) + A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu) + A_{j,\pi}(\bar{h}^{(T)};\nu) + \Gamma_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu) + M_{j,0}(\bar{h}^{(T)};\nu),$$

where

$$A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu) := \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\Big(\mathbb{E}_{\mathcal{P}_{j,\mathbf{Z}|Y=c}}[q_{\nu,c}^{\bar{h}^{(T)}}(\mathbf{Z};\boldsymbol{\phi})] - \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}}[q_{\nu,c}^{\bar{h}^{(T)}}(\mathbf{Z};\boldsymbol{\phi})]\Big),$$

$$A_{j,\pi}(\bar{h}^{(T)};\nu) := \sum_{c\in\mathcal{C}_\nu} (\pi_j(c)-\pi_\nu(c))\mathbb{E}_{\mathcal{P}_\nu}[g_{\bar{h}^{(T)}}(\mathbf{X},Y)\mid Y=c],$$

and

$$\Gamma_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu) := \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\mathbb{E}_{\mathcal{P}_{j,\mathbf{Z}|Y=c}}[q_{j,c}^{\bar{h}^{(T)}}(\mathbf{Z};\boldsymbol{\phi}) - q_{\nu,c}^{\bar{h}^{(T)}}(\mathbf{Z};\boldsymbol{\phi})].$$

The terms $A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu)$ and $\Gamma_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu)$ depend on the conditioning representation. The integrated prior and unsupported-class terms depend only on the predictor, mixture, and target law. By definition,

$$|\Gamma_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu)| = \rho_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu).$$

Step 2: source-mixture robust risk. Since

$$\mathcal{R}_\nu^{\varepsilon}(\bar{h}^{(T)}) = \sum_{i\in S_j} \nu_i \mathcal{R}_i^{\varepsilon}(\bar{h}^{(T)}),$$

the robust risk stability condition gives

$$\mathcal{R}_i^{\varepsilon}(\bar{h}^{(T)}) \le \mathcal{R}_i^{\varepsilon}(h_i^{(T)}) + L_{\mathrm{stab}}\, d_\varepsilon(h_i^{(T)},\bar{h}^{(T)}).$$

Therefore,

$$\mathcal{R}_\nu^{\varepsilon}(\bar{h}^{(T)}) \le L_{\mathrm{src}}(\mathbf{h}^{(T)};\nu) + L_{\mathrm{stab}} \sum_{i\in S_j} \nu_i d_\varepsilon(h_i^{(T)},\bar{h}^{(T)}).$$

By state-to-function stability,

$$d_\varepsilon(h_i^{(T)},\bar{h}^{(T)}) \le M_\varepsilon \|\boldsymbol{\omega}_i^{(T)} - \bar{\boldsymbol{\omega}}^{(T)}\|_2.$$

Using

$$\bar{\boldsymbol{\omega}}^{(T)} = \sum_{k\in S_j} \nu_k \boldsymbol{\omega}_k^{(T)},$$

we obtain

$$\|\boldsymbol{\omega}_i^{(T)} - \bar{\boldsymbol{\omega}}^{(T)}\|_2 \le \|\boldsymbol{\omega}_i^{(T)} - \bar{\boldsymbol{\omega}}^{(T-1)}\|_2 + \|\bar{\boldsymbol{\omega}}^{(T)} - \bar{\boldsymbol{\omega}}^{(T-1)}\|_2 \le \Delta_i^{(T)} + \sum_{k\in S_j} \nu_k \Delta_k^{(T)}.$$

Hence

$$\sum_{i\in S_j} \nu_i d_\varepsilon(h_i^{(T)},\bar{h}^{(T)}) \le 2M_\varepsilon \sum_{i\in S_j} \nu_i \Delta_i^{(T)}.$$

The EMA recursion bounds the trainable-parameter contribution to this drift. If $d_i^{(k)} := \|\boldsymbol{\theta}_i^{(T,k)} - \bar{\boldsymbol{\theta}}^{(T-1)}\|_2$, then

$$d_i^{(k+1)} \le (1-\gamma_i) d_i^{(k)} + (1-\gamma_i)\eta_i G_i,\ d_i^{(0)} = 0,$$

which gives

$$d_i^{(E_i)} \le \eta_i G_i \begin{cases} \frac{(1-\gamma_i)(1-(1-\gamma_i)^{E_i})}{\gamma_i}, & \gamma_i > 0, \\ E_i, & \gamma_i = 0. \end{cases}$$

Hence the EMA tether replaces the untethered linear factor E_i by the smaller geometric factor in the first case, which is the sense in which it controls the trainable-parameter part of the drift. In particular, $\Delta_i^{(T)} \le d_i^{(E_i)} + \|\boldsymbol{b}_i^{(T)} - \bar{\boldsymbol{b}}^{(T-1)}\|_2$; the buffer term is not discounted by parameter tethering.

Thus

$$\mathcal{R}_\nu^{\varepsilon}(\bar{h}^{(T)}) \le L_{\mathrm{src}}(\mathbf{h}^{(T)};\nu) + U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)};\nu).$$

Step 3: class-conditional and prior terms. By discrepancy compatibility,

$$|A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu)| \le C_{\mathrm{cc}}(\boldsymbol{\phi};\nu)\sum_{c\in\mathcal{C}_\nu}\pi_j(c)\,\mathrm{IPM}_{\mathcal{F}}\big(\mathcal{P}_{j,\mathbf{Z}|Y=c},\mathcal{P}_{\nu,\mathbf{Z}|Y=c}\big).$$

The source-coverage condition gives

$$|A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu)| \le C_{\mathrm{cc}}(\boldsymbol{\phi};\nu)\kappa_{\mathrm{cc}}^{j}(\boldsymbol{\phi};\nu)D_{\mathrm{src}}(\boldsymbol{\phi};\nu).$$

Also, since $0 \le q_{\nu,c}^{\bar{h}^{(T)}}(z;\boldsymbol{\phi}) \le B$,

$$|A_{j,\pi}(\bar{h}^{(T)};\nu)| + M_{j,0}(\bar{h}^{(T)};\nu) \le B\|\pi_j - \pi_\nu\|_1.$$

The source-coverage condition also gives

$$|A_{j,\pi}(\bar{h}^{(T)};\nu)| + M_{j,0}(\bar{h}^{(T)};\nu) \le B\kappa_{\pi}^{j}(\nu)r_Y(\nu).$$

Step 4: contrastive calibration of source dispersion. For $C \ge 2$, set $b_{\tau,C} := \log(1 + (C-1)e^{-2/\tau_{\mathrm{con}}})$ and

$$\Psi_{\tau,C}(u) := \min\left\{2,\ 2 + \tau_{\mathrm{con}}\log\frac{e^u - 1}{C-1}\right\}, \quad u \ge b_{\tau,C}.$$

The map $\Psi_{\tau,C}$ is nondecreasing and concave on $[b_{\tau,C},\infty)$: before the clipping point $u = \log C$, it is the inverse of the increasing log-sum-exp lower envelope below, and after that point it is constant. For $\mathbf{Z}, v_c \in \mathbb{S}^{d-1}$ and $Y = c$,

$$\ell_{\mathrm{proto}}(\mathbf{X},c;\boldsymbol{\phi},\mathbf{v}) = \log\left(1 + \sum_{c'\ne c}\exp\left(\frac{\langle\mathbf{Z},v_{c'}\rangle - \langle\mathbf{Z},v_c\rangle}{\tau_{\mathrm{con}}}\right)\right).$$

Let $t := 1 - \langle\mathbf{Z},v_c\rangle \in [0,2]$. Since $\langle\mathbf{Z},v_{c'}\rangle \ge -1$,

$$\ell_{\mathrm{proto}}(\mathbf{X},c;\boldsymbol{\phi},\mathbf{v}) \ge \log\left(1 + (C-1)\exp\left(\frac{t-2}{\tau_{\mathrm{con}}}\right)\right).$$

The right-hand side is increasing in t, and the case $t \le 2$ is already built into the clipping in $\Psi_{\tau,C}$. Hence

$$1 - \langle\mathbf{Z},v_c\rangle = t \le \Psi_{\tau,C}\big(\ell_{\mathrm{proto}}(\mathbf{X},c;\boldsymbol{\phi},\mathbf{v})\big).$$

Using

$$\|\mathbf{Z} - v_c\|_2^2 = 2(1 - \langle\mathbf{Z},v_c\rangle),$$

we get

$$\mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}}\|\mathbf{Z}-v_c\|_2^2 \le 2\,\mathbb{E}_{\mathbf{X}\sim\mathcal{P}_{\nu,\mathbf{X}|Y=c}}\Psi_{\tau,C}\big(\ell_{\mathrm{proto}}(\mathbf{X},c;\boldsymbol{\phi},\mathbf{v})\big) \le 2\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big),$$

where the last step uses Jensen's inequality.

By the centroid imbalance factor,

$$\max_{i\in S_j(c)}\mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}}\|\mathbf{Z}-v_c\|_2^2 \le 2\xi_c(\boldsymbol{\phi},\mathbf{v};\nu)\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big).$$

Therefore,

$$\max_{i\in S_j(c)}\mathbb{E}_{\mathcal{P}_{i,\mathbf{Z}|Y=c}}\|\mathbf{Z}-v_c\|_2 \le \sqrt{2\xi_c(\boldsymbol{\phi},\mathbf{v};\nu)}\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

By Lemma 2,

$$\max_{i\in S_j(c)}\mathrm{IPM}_{\mathcal{F}}\big(\mathcal{P}_{i,\mathbf{Z}|Y=c},\mathcal{P}_{\nu,\mathbf{Z}|Y=c}\big) \le 2\sqrt{2}L_{\mathcal{F}}\sqrt{\xi_c(\boldsymbol{\phi},\mathbf{v};\nu)}\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

Let $\xi_{\max}(\boldsymbol{\phi},\mathbf{v};\nu) := \max_{c\in\mathcal{C}_\nu}\xi_c(\boldsymbol{\phi},\mathbf{v};\nu)$. Then

$$D_{\mathrm{src}}(\boldsymbol{\phi};\nu) \le 2\sqrt{2}L_{\mathcal{F}}\sqrt{\xi_{\max}(\boldsymbol{\phi},\mathbf{v};\nu)}\sum_{c\in\mathcal{C}_\nu}\pi_\nu(c)\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

Since $u \mapsto \sqrt{\Psi_{\tau,C}(u)}$ is concave, Jensen's inequality and the definition of the population prototype loss give

$$\sum_{c\in\mathcal{C}_\nu} \pi_\nu(c)\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto},c}(\boldsymbol{\phi},\mathbf{v};\nu)\big)} \le \sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto}}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

Hence

$$D_{\mathrm{src}}(\boldsymbol{\phi};\nu) \le 2\sqrt{2}L_{\mathcal{F}}\sqrt{\xi_{\max}(\boldsymbol{\phi},\mathbf{v};\nu)}\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto}}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

Set

$$\Lambda^j_{\mathrm{cov}}(\boldsymbol{\phi},\mathbf{v};\nu) := 2\sqrt{2}\,C_{\mathrm{cc}}(\boldsymbol{\phi};\nu)\kappa^j_{\mathrm{cc}}(\boldsymbol{\phi};\nu)L_{\mathcal{F}}\sqrt{\xi_{\max}(\boldsymbol{\phi},\mathbf{v};\nu)}.$$

Here $C_{\mathrm{cc}}(\boldsymbol{\phi};\nu)$ controls the conditional robust-loss profiles for the fixed attack/loss class, $\kappa^j_{\mathrm{cc}}(\boldsymbol{\phi};\nu)$ compares target and source feature geometry, and $\xi_{\max}(\boldsymbol{\phi},\mathbf{v};\nu)$ measures source dispersion around the anchors. The critic class and its constant $L_{\mathcal{F}}$ are fixed. Combining this with Step 3 gives

$$|A_{j,\mathrm{cc}}(\bar{h}^{(T)};\boldsymbol{\phi},\nu)| \le \Lambda^j_{\mathrm{cov}}(\boldsymbol{\phi},\mathbf{v};\nu)\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto}}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}.$$

Combining Steps 1–4 yields the population inequality

$$\begin{aligned}\mathcal{R}^{\varepsilon}_j(\bar{h}^{(T)}) \le\ & L_{\mathrm{src}}(\mathbf{h}^{(T)};\nu) + U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)};\nu)\\ &+ \Lambda^j_{\mathrm{cov}}(\boldsymbol{\phi},\mathbf{v};\nu)\sqrt{\Psi_{\tau,C}\big(L_{\mathrm{proto}}(\boldsymbol{\phi},\mathbf{v};\nu)\big)}\\ &+ B\kappa^j_{\pi}(\nu)r_Y(\nu) + \rho_j(\bar{h}^{(T)};\boldsymbol{\phi},\nu).\end{aligned}$$

On $\mathcal{A}_\beta$, substitute the uniform generalization bounds in Equations (15) to (17). Monotonicity of $\Psi_{\tau,C}$ then gives Equation (3).

This proves the theorem. □

A.5 Conditional Residual and Source Mixtures

Fix $h, \boldsymbol{\phi}, \nu$. We use the input class-conditionals $\mathcal{P}_{r,c} = \mathcal{P}_{r,\mathbf{X}|Y=c}$ from the main text. Only classes $c \in \mathcal{C}_\nu$ with $\pi_j(c) > 0$ contribute to the residual; terms with zero target mass are omitted. We do not assume target label-support inclusion. Classes outside $\mathcal{C}_\nu$ are handled by $M_{j,0}$ and the label-prior term in Appendix A.4.

Proof of Lemma 1. For a contributing class c, define

$$f_c(\boldsymbol{x}) = g_h(\boldsymbol{x},c), \qquad r_c(\boldsymbol{x}) = f_c(\boldsymbol{x}) - q^h_{\nu,c}(\mathbf{Z}_{\boldsymbol{\phi}}(\boldsymbol{x});\boldsymbol{\phi}).$$

The fixed measurable source version makes r_c measurable. Both f_c and this version take values in $[0,B]$, so $|r_c| \le B$ at every input, including those mapped to source-null feature sets. The tower property under the source law gives

$$\mathbb{E}_{\mathcal{P}_{\nu,c}} r_c(\mathbf{X}) = \mathbb{E}_{\mathcal{P}_{\nu,c}} f_c(\mathbf{X}) - \mathbb{E}_{\mathcal{P}_{\nu,\mathbf{Z}|Y=c}} q^h_{\nu,c}(\mathbf{Z};\boldsymbol{\phi}) = 0.$$

Applying the tower property separately under the target law yields

$$\begin{aligned}\Delta_c &:= \mathbb{E}_{\mathcal{P}_{j,\mathbf{Z}|Y=c}}[q^h_{j,c}(\mathbf{Z};\boldsymbol{\phi}) - q^h_{\nu,c}(\mathbf{Z};\boldsymbol{\phi})]\\ &= \mathbb{E}_{\mathcal{P}_{j,c}} r_c(\mathbf{X}) = \int r_c\, d(\mathcal{P}_{j,c} - \mathcal{P}_{\nu,c}).\end{aligned}$$

The two conditional profiles are defined under different laws; no equality between them is used. With $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$, the total variation norm of the signed measure $P - Q$ is $2d_{\mathrm{TV}}(P,Q)$. Thus

$$|\Delta_c| \le 2\|r_c\|_\infty d_{\mathrm{TV}}(\mathcal{P}_{j,c},\mathcal{P}_{\nu,c}) \le 2B\, d_{\mathrm{TV}}(\mathcal{P}_{j,c},\mathcal{P}_{\nu,c}).$$

Finally, $\rho_j(h;\boldsymbol{\phi},\nu) = |\sum_c \pi_j(c)\Delta_c| \le \sum_c \pi_j(c)|\Delta_c|$, giving Equation (4). This argument does not require absolute continuity. It holds for each fixed source version satisfying the stated $[0,B]$ condition, without asserting that the residual is independent of the version on null sets. □

Proof of Corollary 1. For $c \in \mathcal{C}_\nu$ and any measurable set A, conditioning the joint source mixture gives

$$\mathcal{P}_{\nu,c}(A) = \frac{\sum_{i\in S_j} \nu_i \mathcal{P}_i(\mathbf{X} \in A, Y = c)}{\pi_\nu(c)} = \sum_{i\in S_j(c)} \frac{\nu_i \pi_i(c)}{\pi_\nu(c)} \mathcal{P}_{i,c}(A),$$

giving the class-conditional source mixture. The weights $\nu_{i,c}$ sum to one. For any $\alpha_c \in \Delta(S_j(c))$, set $a_i = \alpha_{i,c} - \nu_{i,c}$; then $\sum_i a_i = 0$ and $\sum_i (a_i)_+ = \sum_i (-a_i)_+ = \frac{1}{2}\sum_i |a_i|$. Since $0 \le \mathcal{P}_{i,c}(A) \le 1$,

$$\begin{aligned} d_{\mathrm{TV}}(Q_{\alpha,c}, \mathcal{P}_{\nu,c}) &= \sup_A \left| \sum_{i\in S_j(c)} a_i \mathcal{P}_{i,c}(A) \right| \\ &\le \frac{1}{2} \|\alpha_c - \nu_c\|_1. \end{aligned} \tag{18}$$

The TV triangle inequality and Equation (18) give

$$d_{\mathrm{TV}}(\mathcal{P}_{j,c}, \mathcal{P}_{\nu,c}) \le d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c}) + \tfrac{1}{2}\|\alpha_c - \nu_c\|_1.$$

Substitution into Equation (4) proves Equation (5). When $\mathcal{P}_{j,c} = Q_{\alpha,c}$, its first term vanishes. If also $\alpha_c = \nu_c$ for every contributing class, both terms vanish and nonnegativity gives $\rho_j(h; \boldsymbol{\phi}, \nu) = 0$. Thus matching $\mathcal{P}_{j,c} = \mathcal{P}_{\nu,c}$ is sufficient; membership in the source conditional hull alone is not. □

Distance to the source conditional hull. For each contributing class, define

$$\eta_c := \inf_{\alpha_c \in \Delta(S_j(c))} d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c}).$$

The finite-dimensional simplex is nonempty and compact. The proof of Equation (18), applied to any two weight vectors, shows that

$$|d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c}) - d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha',c})| \le \tfrac{1}{2}\|\alpha_c - \alpha'_c\|_1.$$

The objective is therefore continuous, so the infimum defining η_c is attained. For any minimizer $\alpha_c^\star$, Equation (5) yields

$$\rho_j(h; \boldsymbol{\phi}, \nu) \le 2B \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\eta_c + B \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\|\alpha_c^\star - \nu_c\|_1.$$

The mismatch uses the same minimizer as the hull-approximation term; the terms are not minimized independently. Neither η_c nor $\alpha_c^\star$ is estimated from target data by FedR^{2}DG. The weights ν_c depend on both the aggregation weights and source class priors. Their mismatch with α_c is distinct from the label-prior mismatch $\|\pi_j - \pi_\nu\|_1$.

Exact hull and matched-mixture cases. If $\eta_c = 0$ for every contributing class, only mixture mismatch remains. This need not be zero: with one class of probability one, constant features, source laws concentrated at $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, and $g_h(\boldsymbol{x}_0, c) = 0$, $g_h(\boldsymbol{x}_1, c) = B$, source weights $(1/2, 1/2)$ and target weights $(4/5, 1/5)$ give $\rho_j(h; \boldsymbol{\phi}, \nu) = 3B/10$ despite exact hull membership. If instead $\mathcal{P}_{j,c} = \mathcal{P}_{\nu,c}$ for every contributing class, the residual is zero. These are sufficient conditions; they are not necessary, since the signed class contributions can also cancel.

Corollary 2. *Under the assumptions of Theorem 1 and the source-version condition in Lemma 1, on the same event $\mathcal{A}_\beta$, for any $\alpha_c \in \Delta(S_j(c))$,*

$$\begin{aligned} \mathcal{R}_j^\varepsilon(\bar{h}^{(T)}) \le\ & \widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)}; \nu) + \mathcal{E}_{\mathrm{src}}(\beta; \nu) + U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)}; \nu) \\ & + \Lambda^j_{\mathrm{cov}}(\boldsymbol{\phi}, \mathbf{v}; \nu) \sqrt{\Psi_{\tau,C}\big(\widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi}, \mathbf{v}; \nu) + \mathcal{E}_{\mathrm{proto}}(\beta; \nu)\big)} \\ & + B\kappa^j_\pi(\nu)\big(\widehat{r}_Y(\nu) + \mathcal{E}_{\mathrm{prior}}(\beta)\big) \\ & + 2B \sum_{c\in\mathcal{C}_\nu} \pi_j(c) d_{\mathrm{TV}}(\mathcal{P}_{j,c}, Q_{\alpha,c}) + B \sum_{c\in\mathcal{C}_\nu} \pi_j(c)\|\alpha_c - \nu_c\|_1. \end{aligned}$$

Remark 5. The residual inequality is deterministic, so the bound holds on the same event with probability at least $1 - 3\beta$. No additional sampling event or target-data estimate is used.

Proof. Substitute Equation (5), evaluated at $h = \bar{h}^{(T)}$ and its feature extractor $\boldsymbol{\phi}$, into Equation (3). □

B PROPOSED ALGORITHM

Algorithm 1 summarizes the proposed training procedure. The method consists of three phases: clean federated warm-up, global centroid initialization, and selective-defense federated robust training. During the main training phase, defended clients perform PGD-based adversarial training, while weakly defended clients use a budget-matched replay-based perturbation scheme combined with clean supervision and prototype alignment.

Algorithm 1 FedR^{2}DG: source-only selective robust training

Input: source samples $\{\widehat{D}_i\}_{i\in S_j}$; partition $S_j = \mathcal{D}_j \cup \mathcal{N}_j$; initial model state $(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{b}})$.
Configuration: $T_{\mathrm{warm}}, T, E, B_C, B_H, \zeta$ with $B_C + B_H\zeta = E$; $\varepsilon, \eta_i, \lambda, \tau_{\mathrm{con}}$; $\gamma_{\mathrm{ema}}, \alpha_{\mathrm{ema}}, I_{\mathrm{cent}}$.
Update convention: $\mathrm{Step}_i(\boldsymbol{\theta}, \mathcal{L})$ performs one optimizer step (SGD with momentum and weight decay in the experiments).

Run clean FedAvg for T_{warm} rounds, with E local steps per round.
Broadcast the post-warm-up model; initialize $\{v_c\}$ by Equation (6).
Set $\nu_i = \sqrt{n_i}/\sum_{i'\in S_j}\sqrt{n_{i'}}$ and call the warm-up output $(\bar{\boldsymbol{\theta}}^{(0)}, \bar{\boldsymbol{b}}^{(0)})$.
for $t = 1, \ldots, T$ **do**
  Broadcast $(\bar{\boldsymbol{\theta}}^{(t-1)}, \bar{\boldsymbol{b}}^{(t-1)}, \mathbf{v})$ to all source clients.
  for each source client i **do**
    Initialize $(\boldsymbol{\theta}_i, \boldsymbol{b}_i) \leftarrow (\bar{\boldsymbol{\theta}}^{(t-1)}, \bar{\boldsymbol{b}}^{(t-1)})$.
    if $i \in \mathcal{D}_j$ **then**
      for $k = 1, \ldots, E$ **do**
        Sample a mini-batch and generate $\boldsymbol{x}_{\mathrm{adv}}$ with PGD.
        $\boldsymbol{\theta}_i \leftarrow \mathrm{Step}_i(\boldsymbol{\theta}_i, \mathcal{L}_{\mathrm{def}})$ using Equation (9).
        $\boldsymbol{\theta}_i \leftarrow (1-\gamma_{\mathrm{ema}})\boldsymbol{\theta}_i + \gamma_{\mathrm{ema}}\bar{\boldsymbol{\theta}}^{(t-1)}$.
      end for
    else
      for $k = 1, \ldots, B_C$ **do**
        Sample a mini-batch; $\boldsymbol{\theta}_i \leftarrow \mathrm{Step}_i(\boldsymbol{\theta}_i, \mathcal{L}_{\mathrm{clean}})$ using Equation (10).
        Tether $\boldsymbol{\theta}_i$ to $\bar{\boldsymbol{\theta}}^{(t-1)}$ as in line 12.
      end for
      for $k = 1, \ldots, B_H$ **do**
        Sample one mini-batch and reset $\boldsymbol{\delta}^{(0)} \leftarrow 0$.
        for $r = 1, \ldots, \zeta$ **do**
          Compute $\mathcal{L}_{\mathrm{rep}}$ and both parameter and input gradients in one backward pass.
          Apply Step_i using the same parameter gradient, then tether as in line 12.
          Update $\boldsymbol{\delta}^{(r)}$ by Equation (11) using the current input gradient.
        end for
      end for
    end if
    Store $(\boldsymbol{\theta}_i^{(t)}, \boldsymbol{b}_i^{(t)}) \leftarrow (\boldsymbol{\theta}_i, \boldsymbol{b}_i)$.
    if $t \bmod I_{\mathrm{cent}} = 0$ **then**
      Compute $\{v_{i,c}\}$ in evaluation mode using the updated local model.
    end if
  end for
  Aggregate parameters and prediction buffers by Equation (12).
  if $t \bmod I_{\mathrm{cent}} = 0$ **then**
    Form $\widehat{v}_c$ by class-count weighting; update v_c by Equation (7).
  end if
end for
return $(\bar{\boldsymbol{\theta}}^{(T)}, \bar{\boldsymbol{b}}^{(T)})$.

Table 4 connects the algorithmic components to the bound.

Table 4: Algorithmic components and the corresponding bound signals in Theorem 1 that they directly target or indirectly influence.

Component	**Bound signal**	**Algorithmic role**
PGD-AT	$\widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)};\nu)$	Robust source fitting
Budgeted replay	$\widehat{L}_{\mathrm{src}}(\mathbf{h}^{(T)};\nu)$	Weak client robust proxy
EMA tethering	$U_{\mathrm{drift}}(\boldsymbol{\Delta}^{(T)};\nu)$	Parameter-drift control
Prototype loss	$\widehat{L}_{\mathrm{proto}}(\boldsymbol{\phi},\mathbf{v};\nu)$	Class-conditional alignment
Geometric-mean weights	$\widehat{r}_Y(\nu)$	Source-mixture rebalancing

C ROBUSTNESS AFTER CLEAN FINE-TUNING

This section reports the robustness of models trained without any task-specific adversarial training. All clients are trained only with clean cross-entropy, and robustness is evaluated post hoc against AA. These results serve as a diagnostic baseline for measuring how much adversarial robustness can be retained without explicit robust optimization on the source datasets. All models start from the same adversarially pretrained ImageNet backbone as the main experiments; these are clean fine-tuning baselines, not models trained without any prior robust pretraining.

C.1 PACS

Table 5: All-clean CE baseline results. ERM is trained separately on each source client, and the results are averaged over source clients. Standard accuracy is reported once because it is independent of the attack budget. The ± values denote standard errors over three seeds.

Method	Metric	ε	Target A	Target C	Target P	Target S
ERM	Standard.	–	$56.09_{\pm 0.87}$	$48.72_{\pm 2.03}$	$72.80_{\pm 0.91}$	$55.95_{\pm 1.49}$
	Rob.	$2/255$	$22.08_{\pm 0.62}$	$28.60_{\pm 0.99}$	$46.72_{\pm 1.04}$	$47.79_{\pm 1.49}$
		$4/255$	$3.69_{\pm 0.41}$	$14.66_{\pm 1.19}$	$16.90_{\pm 2.16}$	$37.93_{\pm 2.09}$
		$8/255$	$0.06_{\pm 0.02}$	$2.86_{\pm 0.40}$	$0.45_{\pm 0.12}$	$13.84_{\pm 2.94}$
FedAvg	Standard.	–	$80.65_{\pm 0.53}$	$80.56_{\pm 0.45}$	$93.81_{\pm 0.36}$	$84.68_{\pm 0.96}$
	Rob.	$2/255$	$35.54_{\pm 1.09}$	$56.68_{\pm 0.89}$	$68.18_{\pm 0.43}$	$73.56_{\pm 1.08}$
		$4/255$	$7.16_{\pm 0.19}$	$31.47_{\pm 0.91}$	$28.58_{\pm 0.54}$	$55.42_{\pm 0.51}$
		$8/255$	$0.10_{\pm 0.00}$	$4.76_{\pm 0.52}$	$0.46_{\pm 0.02}$	$6.25_{\pm 0.46}$
FedProx	Standard.	–	$79.88_{\pm 1.30}$	$80.29_{\pm 0.32}$	$93.47_{\pm 0.40}$	$84.75_{\pm 1.11}$
	Rob.	$2/255$	$34.23_{\pm 1.34}$	$56.75_{\pm 0.91}$	$67.83_{\pm 0.76}$	$73.44_{\pm 0.95}$
		$4/255$	$6.05_{\pm 0.32}$	$32.13_{\pm 0.82}$	$27.58_{\pm 1.20}$	$54.58_{\pm 1.09}$
		$8/255$	$0.08_{\pm 0.02}$	$5.02_{\pm 0.32}$	$0.44_{\pm 0.09}$	$6.81_{\pm 0.98}$

C.2 OFFICEHOME

Table 6: All-clean CE baseline results. ERM is trained separately on each source client, and the results are averaged over source clients. Standard accuracy is reported once because it is independent of the attack budget. The ± values denote standard errors over three seeds.

Method	Metric	ε	Target Ar	Target Cl	Target Rw	Target Pr
ERM	Standard.	–	$50.83_{\pm 0.30}$	$48.34_{\pm 0.37}$	$65.77_{\pm 0.27}$	$63.76_{\pm 0.29}$
	Rob.	$2/255$	$12.86_{\pm 0.35}$	$23.63_{\pm 0.60}$	$24.77_{\pm 0.56}$	$28.85_{\pm 0.83}$
		$4/255$	$2.21_{\pm 0.13}$	$6.69_{\pm 0.38}$	$5.13_{\pm 0.18}$	$6.67_{\pm 0.37}$
		$8/255$	$0.03_{\pm 0.01}$	$0.54_{\pm 0.05}$	$0.30_{\pm 0.04}$	$0.54_{\pm 0.07}$
FedAvg	Standard.	–	$65.14_{\pm 0.30}$	$58.41_{\pm 0.31}$	$77.98_{\pm 0.01}$	$76.24_{\pm 0.07}$
	Rob.	$2/255$	$10.60_{\pm 0.45}$	$21.59_{\pm 0.50}$	$21.57_{\pm 0.89}$	$27.28_{\pm 0.40}$
		$4/255$	$0.85_{\pm 0.06}$	$3.74_{\pm 0.30}$	$2.57_{\pm 0.20}$	$3.30_{\pm 0.10}$
		$8/255$	$0.04_{\pm 0.00}$	$0.09_{\pm 0.01}$	$0.10_{\pm 0.01}$	$0.18_{\pm 0.05}$
FedProx	Standard.	–	$65.34_{\pm 0.34}$	$58.06_{\pm 0.39}$	$77.91_{\pm 0.07}$	$76.14_{\pm 0.15}$
	Rob.	$2/255$	$10.65_{\pm 0.61}$	$21.64_{\pm 0.46}$	$21.71_{\pm 0.17}$	$28.30_{\pm 0.98}$
		$4/255$	$0.84_{\pm 0.11}$	$3.59_{\pm 0.24}$	$2.65_{\pm 0.10}$	$3.41_{\pm 0.07}$
		$8/255$	$0.04_{\pm 0.00}$	$0.07_{\pm 0.01}$	$0.12_{\pm 0.01}$	$0.15_{\pm 0.01}$

C.3 MINI-DOMAINNET50

Table 7: All-clean CE baseline results. ERM is trained separately on each source client, and the results are averaged over source clients. Standard accuracy is reported once because it is independent of the attack budget. The $\pm$ values denote standard errors over three seeds.

Method	Metric	ε	Target Clipart	Target Painting	Target Real	Target Sketch
ERM	Standard.	–	$57.46_{\pm 0.69}$	$57.69_{\pm 0.54}$	$68.71_{\pm 0.56}$	$56.74_{\pm 0.50}$
	Rob.	$2/255$	$28.53_{\pm 0.43}$	$12.64_{\pm 0.28}$	$23.54_{\pm 0.66}$	$23.73_{\pm 0.85}$
		$4/255$	$9.33_{\pm 0.49}$	$1.64_{\pm 0.08}$	$3.91_{\pm 0.23}$	$7.84_{\pm 0.82}$
		$8/255$	$0.81_{\pm 0.09}$	$0.07_{\pm 0.01}$	$0.18_{\pm 0.02}$	$0.53_{\pm 0.07}$
FedAvg	Standard.	–	$74.37_{\pm 0.19}$	$73.21_{\pm 0.17}$	$81.11_{\pm 0.25}$	$72.16_{\pm 0.30}$
	Rob.	$2/255$	$33.21_{\pm 0.33}$	$11.11_{\pm 0.47}$	$21.15_{\pm 0.41}$	$25.93_{\pm 0.91}$
		$4/255$	$6.92_{\pm 0.12}$	$0.74_{\pm 0.04}$	$2.05_{\pm 0.03}$	$5.55_{\pm 0.47}$
		$8/255$	$0.28_{\pm 0.02}$	$0.02_{\pm 0.00}$	$0.07_{\pm 0.01}$	$0.17_{\pm 0.02}$
FedProx	Standard.	–	$74.72_{\pm 0.18}$	$73.31_{\pm 0.12}$	$80.90_{\pm 0.17}$	$72.43_{\pm 0.44}$
	Rob.	$2/255$	$35.05_{\pm 1.07}$	$11.14_{\pm 0.74}$	$20.99_{\pm 0.52}$	$26.59_{\pm 0.62}$
		$4/255$	$8.15_{\pm 0.99}$	$0.67_{\pm 0.08}$	$1.89_{\pm 0.08}$	$5.85_{\pm 0.61}$
		$8/255$	$0.23_{\pm 0.02}$	$0.02_{\pm 0.00}$	$0.09_{\pm 0.02}$	$0.19_{\pm 0.07}$

D Detailed Per-Case Results

We report means and standard errors over three seeds. Aggregate values first average the 12 target/defended-client cases within each seed. The harmonic mean is computed for each case and seed before aggregation, so it need not equal the harmonic mean of the displayed aggregate accuracies. Bold and underlined entries indicate the best and second-best means, including ties.

D.1 Aggregate Results with Uncertainty

Table 8: PACS: aggregate means ± standard errors (%).

ε	Method	Standard	Robust	H
2/255	Robust ERM	63.93 ± 0.44	49.24 ± 0.25	55.07 ± 0.33
	SR-FedAvg	85.25 ± 0.20	66.19 ± 0.42	73.98 ± 0.37
	SR-FRP	84.89 ± 0.20	<u>66.95</u> ± 0.32	<u>74.29</u> ± 0.25
	SR-FedProx	<u>85.27</u> ± 0.26	66.20 ± 0.43	73.99 ± 0.38
	FedR^{2}DG	**85.41** ± 0.24	**72.64** ± 0.25	**78.34** ± 0.25
4/255	Robust ERM	62.99 ± 0.62	40.14 ± 0.27	47.44 ± 0.32
	SR-FedAvg	83.92 ± 0.14	50.74 ± 0.36	60.93 ± 0.39
	SR-FRP	**84.21** ± 0.20	47.57 ± 0.48	57.96 ± 0.45
	SR-FedProx	<u>84.00</u> ± 0.19	<u>50.81</u> ± 0.34	<u>61.01</u> ± 0.34
	FedR^{2}DG	83.94 ± 0.31	**61.38** ± 0.29	**70.12** ± 0.26
8/255	Robust ERM	60.91 ± 0.37	29.71 ± 0.23	36.68 ± 0.15
	SR-FedAvg	80.83 ± 0.18	32.45 ± 0.22	41.32 ± 0.22
	SR-FRP	**82.61** ± 0.33	24.22 ± 0.27	31.22 ± 0.30
	SR-FedProx	80.79 ± 0.21	<u>32.50</u> ± 0.24	<u>41.38</u> ± 0.26
	FedR^{2}DG	<u>80.90</u> ± 0.21	**44.49** ± 0.38	**54.64** ± 0.35

Table 9: OfficeHome: aggregate means ± standard errors (%).

ε	Method	Standard	Robust	H
2/255	Robust ERM	56.06 ± 0.13	41.33 ± 0.20	47.38 ± 0.19
	SR-FedAvg	<u>68.77</u> ± 0.10	46.98 ± 0.07	55.45 ± 0.05
	SR-FRP	**68.98** ± 0.05	42.69 ± 0.31	52.13 ± 0.26
	SR-FedProx	68.75 ± 0.13	<u>47.04</u> ± 0.06	<u>55.49</u> ± 0.06
	FedR^{2}DG	68.27 ± 0.04	**51.95** ± 0.04	**58.80** ± 0.05
4/255	Robust ERM	53.63 ± 0.03	32.42 ± 0.11	39.99 ± 0.08
	SR-FedAvg	<u>67.28</u> ± 0.01	<u>34.86</u> ± 0.19	<u>45.16</u> ± 0.18
	SR-FRP	**68.39** ± 0.08	22.96 ± 0.18	33.33 ± 0.26
	SR-FedProx	67.25 ± 0.07	<u>34.86</u> ± 0.18	45.13 ± 0.18
	FedR^{2}DG	66.89 ± 0.05	**41.82** ± 0.04	**50.99** ± 0.02
8/255	Robust ERM	48.31 ± 0.12	<u>21.77</u> ± 0.06	29.43 ± 0.06
	SR-FedAvg	63.94 ± 0.02	20.76 ± 0.14	30.25 ± 0.14
	SR-FRP	**67.01** ± 0.05	6.19 ± 0.15	10.84 ± 0.22
	SR-FedProx	63.99 ± 0.02	20.88 ± 0.14	<u>30.37</u> ± 0.19
	FedR^{2}DG	<u>64.06</u> ± 0.04	**28.66** ± 0.08	**38.84** ± 0.08

Table 10: mini-DomainNet50: aggregate means ± standard errors (%).

ε	Method	Standard	Robust	H
2/255	Robust ERM	57.65 ± 0.14	41.04 ± 0.10	47.78 ± 0.11
	SR-FedAvg	73.86 ± 0.10	49.03 ± 0.12	58.59 ± 0.12
	SR-FRP	**74.61** ± 0.15	44.86 ± 0.29	55.40 ± 0.32
	SR-FedProx	73.85 ± 0.11	49.04 ± 0.07	58.59 ± 0.08
	FedR^{2}DG	73.40 ± 0.07	**55.23** ± 0.04	**62.88** ± 0.03
4/255	Robust ERM	53.63 ± 0.19	30.78 ± 0.12	38.73 ± 0.14
	SR-FedAvg	71.71 ± 0.02	34.14 ± 0.19	45.30 ± 0.19
	SR-FRP	**73.79** ± 0.16	24.23 ± 0.38	35.00 ± 0.37
	SR-FedProx	71.73 ± 0.11	34.03 ± 0.14	45.18 ± 0.17
	FedR^{2}DG	71.58 ± 0.19	**42.85** ± 0.04	**53.12** ± 0.07
8/255	Robust ERM	46.49 ± 0.45	19.45 ± 0.27	26.80 ± 0.34
	SR-FedAvg	67.03 ± 0.22	18.88 ± 0.11	27.95 ± 0.18
	SR-FRP	**72.17** ± 0.05	7.53 ± 0.05	12.89 ± 0.08
	SR-FedProx	66.82 ± 0.32	18.80 ± 0.18	27.94 ± 0.17
	FedR^{2}DG	67.74 ± 0.19	**27.34** ± 0.03	**37.73** ± 0.05

D.2 PACS

Domain codes: A = Art Painting, C = Cartoon, P = Photo, S = Sketch.

Table 11: PACS, target A: per-case accuracy (%), mean ± standard error.

		Defended C		Defended P		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	76.95 ± 0.38	50.26 ± 0.07	71.68 ± 0.22	46.01 ± 0.17	46.21 ± 2.65	19.66 ± 1.79
	SR-FedAvg	81.89 ± 0.42	50.49 ± 0.98	80.29 ± 0.29	47.02 ± 0.75	82.80 ± 0.73	42.97 ± 1.28
	SR-FRP	79.25 ± 0.68	49.92 ± 0.48	77.00 ± 1.06	40.72 ± 0.88	**83.38** ± 0.47	46.04 ± 0.48
	SR-FedProx	81.82 ± 0.59	50.49 ± 0.82	81.07 ± 0.84	47.56 ± 1.24	82.71 ± 0.75	43.02 ± 0.72
	FedR^{2}DG	**83.19** ± 0.95	**60.53** ± 0.72	**83.92** ± 0.42	**60.06** ± 0.25	82.57 ± 1.41	**59.83** ± 1.02
4/255	Robust ERM	71.19 ± 0.82	28.94 ± 0.38	69.13 ± 0.74	29.98 ± 0.40	45.69 ± 3.48	7.93 ± 0.45
	SR-FedAvg	80.99 ± 0.71	28.74 ± 0.41	80.32 ± 0.59	27.03 ± 0.30	82.19 ± 0.54	12.86 ± 0.44
	SR-FRP	78.68 ± 0.33	22.45 ± 0.52	77.52 ± 0.79	13.20 ± 0.41	**83.89** ± 0.47	14.45 ± 0.42
	SR-FedProx	80.96 ± 0.62	28.30 ± 0.29	80.11 ± 0.94	26.59 ± 0.21	82.57 ± 0.66	12.48 ± 0.31
	FedR^{2}DG	**82.47** ± 1.06	**40.09** ± 0.34	**81.77** ± 0.23	**41.05** ± 0.19	81.28 ± 1.25	**38.67** ± 0.49
8/255	Robust ERM	60.87 ± 0.81	10.71 ± 0.64	61.75 ± 0.43	13.09 ± 0.33	47.87 ± 1.20	0.62 ± 0.13
	SR-FedAvg	75.41 ± 0.97	8.09 ± 0.31	77.57 ± 0.39	9.02 ± 0.23	81.20 ± 1.04	0.32 ± 0.03
	SR-FRP	75.55 ± 0.36	2.47 ± 0.32	74.20 ± 0.97	0.83 ± 0.15	**83.76** ± 0.75	0.31 ± 0.04
	SR-FedProx	74.43 ± 0.75	8.10 ± 0.37	77.65 ± 0.44	8.79 ± 0.10	81.44 ± 1.08	0.31 ± 0.04
	FedR^{2}DG	**78.70** ± 0.64	**15.04** ± 0.46	**79.00** ± 0.24	**17.40** ± 0.31	80.00 ± 0.63	**12.27** ± 0.38

Table 12: PACS, target C: per-case accuracy (%), mean ± standard error.

		Defended A		Defended P		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	69.54 ± 1.23	59.11 ± 0.56	42.56 ± 0.82	32.72 ± 0.37	68.18 ± 0.30	53.11 ± 0.41
	SR-FedAvg	80.63 ± 0.97	66.97 ± 1.22	80.43 ± 0.43	61.32 ± 0.63	82.41 ± 0.95	64.89 ± 1.27
	SR-FRP	81.87 ± 0.81	68.20 ± 1.49	80.40 ± 0.79	64.08 ± 0.99	81.94 ± 0.43	66.17 ± 0.38
	SR-FedProx	80.94 ± 0.90	67.22 ± 1.10	80.00 ± 0.64	61.31 ± 1.02	82.31 ± 0.68	64.93 ± 0.78
	FedR^{2}DG	**82.62** ± 0.07	**71.37** ± 0.26	**80.60** ± 0.34	**68.60** ± 0.56	**82.60** ± 0.22	**71.25** ± 0.15
4/255	Robust ERM	70.15 ± 1.24	53.89 ± 0.84	44.89 ± 1.37	29.39 ± 1.95	70.51 ± 0.27	43.97 ± 0.35
	SR-FedAvg	79.39 ± 0.75	57.47 ± 0.84	78.58 ± 0.60	48.18 ± 0.73	**82.18** ± 0.74	50.28 ± 0.78
	SR-FRP	81.23 ± 1.04	55.94 ± 1.74	**79.79** ± 0.67	48.58 ± 0.90	81.64 ± 0.29	50.33 ± 0.36
	SR-FedProx	79.58 ± 0.65	56.98 ± 0.92	78.73 ± 0.61	48.38 ± 0.78	82.00 ± 0.43	50.37 ± 0.85
	FedR^{2}DG	**81.27** ± 0.35	**62.91** ± 0.46	79.00 ± 0.39	**60.11** ± 0.45	81.25 ± 0.15	**61.58** ± 0.50
8/255	Robust ERM	66.94 ± 0.41	43.40 ± 0.26	45.94 ± 0.70	23.56 ± 0.78	71.26 ± 0.42	26.25 ± 0.94
	SR-FedAvg	77.32 ± 0.15	44.10 ± 0.35	76.02 ± 0.69	34.29 ± 0.65	80.86 ± 0.03	25.33 ± 0.52
	SR-FRP	**79.69** ± 0.43	33.29 ± 0.97	**78.16** ± 0.30	24.37 ± 0.36	**81.16** ± 0.35	22.94 ± 0.39
	SR-FedProx	77.20 ± 0.21	44.37 ± 0.39	76.31 ± 0.84	34.60 ± 0.73	80.93 ± 0.24	25.62 ± 0.59
	FedR^{2}DG	77.52 ± 0.14	**49.57** ± 0.54	75.53 ± 0.89	**46.54** ± 0.78	79.82 ± 0.22	**45.62** ± 0.46

Table 13: PACS, target P: per-case accuracy (%), mean ± standard error.

		Defended A		Defended C		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	**93.79** ± 0.14	**82.02** ± 0.14	87.13 ± 0.92	70.00 ± 1.41	41.89 ± 3.44	20.84 ± 2.23
	SR-FedAvg	91.82 ± 0.25	77.17 ± 0.18	**93.77** ± 0.09	76.59 ± 0.19	92.97 ± 0.24	69.50 ± 0.41
	SR-FRP	91.06 ± 0.25	79.24 ± 0.72	93.59 ± 0.14	78.72 ± 0.83	**94.75** ± 0.29	72.64 ± 0.24
	SR-FedProx	91.84 ± 0.28	77.05 ± 0.02	93.61 ± 0.14	76.51 ± 0.35	93.03 ± 0.11	69.10 ± 0.77
	FedR^{2}DG	92.33 ± 0.44	80.72 ± 0.33	91.30 ± 0.69	**79.36** ± 0.76	89.80 ± 0.27	**78.28** ± 0.11
4/255	Robust ERM	**89.60** ± 0.20	69.00 ± 0.42	82.62 ± 0.83	55.59 ± 1.04	42.93 ± 4.92	11.38 ± 0.99
	SR-FedAvg	88.56 ± 0.39	64.79 ± 0.67	91.70 ± 0.26	60.62 ± 0.26	91.32 ± 0.12	39.28 ± 2.05
	SR-FRP	89.56 ± 0.34	44.61 ± 1.23	**91.78** ± 0.36	55.63 ± 1.37	**94.01** ± 0.12	43.58 ± 1.63
	SR-FedProx	88.76 ± 0.26	64.79 ± 0.33	91.76 ± 0.24	60.62 ± 0.30	91.52 ± 0.11	42.04 ± 1.07
	FedR^{2}DG	88.30 ± 0.29	**69.44** ± 0.27	88.82 ± 0.75	**66.29** ± 1.09	87.39 ± 0.21	**65.27** ± 0.18
8/255	Robust ERM	80.20 ± 0.90	**51.32** ± 0.31	72.00 ± 1.43	35.47 ± 0.75	45.37 ± 2.09	2.04 ± 0.72
	SR-FedAvg	83.47 ± 0.03	45.11 ± 0.14	85.77 ± 0.19	31.56 ± 1.50	88.44 ± 0.89	2.60 ± 0.23
	SR-FRP	**87.80** ± 0.36	5.81 ± 0.42	**88.64** ± 0.63	10.12 ± 0.92	**92.93** ± 0.40	2.30 ± 0.07
	SR-FedProx	83.71 ± 0.19	45.29 ± 0.38	85.79 ± 0.33	31.97 ± 1.16	88.54 ± 0.98	2.75 ± 0.09
	FedR^{2}DG	81.96 ± 0.38	51.10 ± 0.14	84.31 ± 0.78	**45.97** ± 0.41	84.43 ± 0.09	**38.86** ± 2.20

Table 14: PACS, target S: per-case accuracy (%), mean ± standard error.

		Defended A		Defended C		Defended P	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	61.32 ± 0.30	56.38 ± 0.02	77.30 ± 1.63	73.02 ± 1.69	30.58 ± 1.35	27.75 ± 1.14
	SR-FedAvg	84.58 ± 0.53	78.96 ± 0.69	**86.32** ± 0.38	80.34 ± 0.41	85.10 ± 0.74	78.09 ± 0.89
	SR-FRP	84.86 ± 0.55	78.62 ± 0.19	85.75 ± 0.52	80.31 ± 0.35	84.88 ± 0.85	78.77 ± 0.84
	SR-FedProx	84.47 ± 0.38	79.15 ± 0.55	86.26 ± 0.32	80.14 ± 0.21	85.14 ± 0.79	77.97 ± 1.17
	FedR^{2}DG	**84.92** ± 0.35	**79.95** ± 0.36	85.80 ± 0.47	**81.01** ± 0.33	**85.27** ± 0.17	**80.77** ± 0.42
4/255	Robust ERM	60.75 ± 0.88	54.08 ± 0.67	76.68 ± 0.92	70.22 ± 0.82	31.72 ± 2.07	27.33 ± 1.69
	SR-FedAvg	81.63 ± 0.37	72.52 ± 0.25	86.33 ± 0.51	76.40 ± 0.41	83.77 ± 0.32	70.68 ± 0.33
	SR-FRP	82.42 ± 0.39	71.90 ± 1.74	85.76 ± 0.48	76.10 ± 0.17	84.19 ± 1.13	74.07 ± 1.59
	SR-FedProx	81.41 ± 0.28	71.98 ± 0.26	86.46 ± 0.45	76.19 ± 0.42	84.11 ± 0.52	70.95 ± 0.20
	FedR^{2}DG	**84.27** ± 0.18	**76.11** ± 0.38	**86.66** ± 0.26	**78.50** ± 0.56	**84.82** ± 0.38	**76.52** ± 0.12
8/255	Robust ERM	63.51 ± 1.47	53.14 ± 1.15	76.39 ± 0.58	67.23 ± 0.34	38.78 ± 3.54	29.69 ± 2.96
	SR-FedAvg	79.13 ± 0.52	64.74 ± 0.78	86.70 ± 0.57	69.24 ± 0.97	78.10 ± 1.17	54.97 ± 1.51
	SR-FRP	79.53 ± 1.11	64.99 ± 0.92	**87.17** ± 0.29	61.88 ± 2.39	**82.70** ± 1.44	61.28 ± 2.48
	SR-FedProx	78.78 ± 0.55	64.40 ± 1.05	86.65 ± 0.48	69.28 ± 0.76	77.99 ± 1.19	54.53 ± 1.33
	FedR^{2}DG	**82.09** ± 0.63	**70.11** ± 0.52	85.49 ± 0.31	**73.21** ± 0.50	81.97 ± 0.40	**68.16** ± 0.67

D.3 OFFICEHOME

Domain codes: A = Art, C = Clipart, P = Product, R = Real-World.

Table 15: OfficeHome, target A: per-case accuracy (%), mean ± standard error.

		Defended C		Defended R		Defended P	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	45.23 ± 0.19	24.80 ± 0.42	58.70 ± 0.40	**38.59** ± 0.42	41.16 ± 0.26	22.39 ± 0.14
	SR-FedAvg	63.96 ± 0.26	29.29 ± 0.13	63.73 ± 0.24	36.00 ± 0.39	62.89 ± 0.19	31.19 ± 0.27
	SR-FRP	**64.58** ± 0.39	22.09 ± 0.58	**64.51** ± 0.48	29.44 ± 0.78	**63.33** ± 0.10	28.21 ± 0.17
	SR-FedProx	63.71 ± 0.29	29.46 ± 0.24	63.55 ± 0.26	35.86 ± 0.33	63.22 ± 0.21	31.08 ± 0.42
	FedR^{2}DG	61.59 ± 0.32	**37.84** ± 0.13	62.38 ± 0.14	38.40 ± 0.19	61.49 ± 0.48	**36.85** ± 0.11
4/255	Robust ERM	41.74 ± 0.30	15.46 ± 0.38	55.23 ± 0.26	26.22 ± 0.41	37.31 ± 0.11	13.93 ± 0.23
	SR-FedAvg	61.40 ± 0.34	15.09 ± 0.16	60.06 ± 0.33	24.32 ± 0.23	60.14 ± 0.28	18.65 ± 0.04
	SR-FRP	**63.30** ± 0.12	6.64 ± 0.44	**62.11** ± 0.31	13.25 ± 0.88	**62.05** ± 0.30	11.74 ± 0.46
	SR-FedProx	61.45 ± 0.12	14.82 ± 0.36	60.09 ± 0.71	24.09 ± 0.19	60.09 ± 0.29	18.57 ± 0.12
	FedR^{2}DG	59.04 ± 0.13	**24.39** ± 0.17	59.22 ± 0.17	**26.42** ± 0.22	58.59 ± 0.29	**24.35** ± 0.15
8/255	Robust ERM	36.88 ± 0.38	7.73 ± 0.12	45.82 ± 0.47	14.54 ± 0.28	29.72 ± 0.57	7.70 ± 0.15
	SR-FedAvg	57.71 ± 0.49	4.96 ± 0.41	52.75 ± 0.28	13.71 ± 0.29	53.30 ± 0.35	7.76 ± 0.37
	SR-FRP	**61.28** ± 0.15	0.38 ± 0.12	**59.62** ± 0.60	2.25 ± 0.28	**58.60** ± 0.33	2.36 ± 0.14
	SR-FedProx	57.46 ± 0.47	4.85 ± 0.34	53.57 ± 1.04	12.67 ± 1.12	53.23 ± 0.20	8.02 ± 0.23
	FedR^{2}DG	54.84 ± 0.21	**12.17** ± 0.07	53.81 ± 0.05	**15.22** ± 0.13	52.81 ± 0.30	**13.06** ± 0.15

Table 16: OfficeHome, target C: per-case accuracy (%), mean ± standard error.

		Defended A		Defended R		Defended P	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	50.43 ± 0.38	41.45 ± 0.23	54.62 ± 0.46	46.28 ± 0.44	44.60 ± 0.13	36.03 ± 0.30
	SR-FedAvg	59.45 ± 0.30	45.37 ± 0.26	59.98 ± 0.06	48.12 ± 0.15	59.28 ± 0.30	45.44 ± 0.25
	SR-FRP	59.70 ± 0.11	42.18 ± 1.03	59.82 ± 0.24	45.33 ± 0.33	59.09 ± 0.72	44.53 ± 0.26
	SR-FedProx	59.31 ± 0.26	45.32 ± 0.36	59.86 ± 0.02	48.06 ± 0.06	59.52 ± 0.36	45.57 ± 0.36
	FedR^{2}DG	**59.96** ± 0.14	**50.15** ± 0.30	**60.34** ± 0.11	**50.81** ± 0.18	**59.80** ± 0.31	**49.42** ± 0.16
4/255	Robust ERM	49.51 ± 0.34	35.94 ± 0.42	54.27 ± 0.08	41.82 ± 0.17	44.37 ± 0.37	31.09 ± 0.44
	SR-FedAvg	59.48 ± 0.36	37.78 ± 0.14	60.47 ± 0.20	43.00 ± 0.40	59.25 ± 0.40	37.18 ± 0.73
	SR-FRP	59.72 ± 0.36	21.42 ± 1.46	**60.81** ± 0.30	33.30 ± 1.30	59.03 ± 0.33	30.20 ± 1.04
	SR-FedProx	59.40 ± 0.28	37.64 ± 0.38	60.43 ± 0.37	43.15 ± 0.51	59.15 ± 0.24	36.89 ± 0.69
	FedR^{2}DG	**60.48** ± 0.50	**44.82** ± 0.47	60.70 ± 0.19	**45.65** ± 0.08	**60.02** ± 0.25	**44.15** ± 0.31
8/255	Robust ERM	45.60 ± 0.20	27.13 ± 0.17	52.15 ± 0.28	34.15 ± 0.54	42.06 ± 0.29	24.19 ± 0.07
	SR-FedAvg	59.50 ± 0.56	26.80 ± 0.51	59.67 ± 0.50	33.75 ± 0.72	58.33 ± 0.31	26.21 ± 0.16
	SR-FRP	59.78 ± 0.27	4.71 ± 0.28	**60.81** ± 0.36	13.49 ± 0.51	58.69 ± 0.14	13.14 ± 0.85
	SR-FedProx	59.41 ± 0.26	26.86 ± 0.38	59.71 ± 0.33	34.58 ± 0.44	58.53 ± 0.51	26.27 ± 0.16
	FedR^{2}DG	**59.87** ± 0.42	**36.10** ± 0.26	60.19 ± 0.31	**37.82** ± 0.14	**59.56** ± 0.15	**35.24** ± 0.14

Table 17: OfficeHome, target R: per-case accuracy (%), mean ± standard error.

		Defended A		Defended C		Defended P	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	67.83 ± 0.21	51.38 ± 0.31	58.26 ± 0.18	39.75 ± 0.25	66.05 ± 0.19	47.51 ± 0.14
	SR-FedAvg	**77.09** ± 0.03	52.63 ± 0.97	**77.06** ± 0.14	50.07 ± 0.17	77.32 ± 0.17	53.03 ± 0.21
	SR-FRP	77.07 ± 0.02	44.20 ± 1.18	76.63 ± 0.31	43.96 ± 1.45	76.80 ± 0.34	49.39 ± 0.60
	SR-FedProx	76.98 ± 0.01	53.05 ± 0.34	76.77 ± 0.13	50.15 ± 0.21	**77.39** ± 0.25	53.00 ± 0.31
	FedR^{2}DG	77.04 ± 0.09	**58.90** ± 0.17	76.42 ± 0.15	**57.75** ± 0.16	76.75 ± 0.34	**58.27** ± 0.08
4/255	Robust ERM	64.38 ± 0.22	39.50 ± 0.05	56.53 ± 0.13	29.03 ± 0.59	62.17 ± 0.41	35.60 ± 0.26
	SR-FedAvg	74.87 ± 0.25	39.28 ± 0.16	75.81 ± 0.06	31.92 ± 0.64	75.72 ± 0.16	38.72 ± 0.28
	SR-FRP	**76.55** ± 0.13	18.58 ± 0.25	**76.97** ± 0.14	17.54 ± 0.38	**76.12** ± 0.35	29.00 ± 0.98
	SR-FedProx	74.88 ± 0.25	39.59 ± 0.44	75.94 ± 0.05	31.86 ± 0.62	75.52 ± 0.19	38.82 ± 0.24
	FedR^{2}DG	75.41 ± 0.02	**46.18** ± 0.35	74.92 ± 0.06	**44.34** ± 0.13	75.51 ± 0.31	**45.33** ± 0.15
8/255	Robust ERM	57.09 ± 0.21	23.14 ± 0.09	52.48 ± 0.30	16.96 ± 0.24	55.78 ± 0.30	23.11 ± 0.06
	SR-FedAvg	70.35 ± 0.18	21.43 ± 0.46	72.94 ± 0.23	13.70 ± 0.67	71.47 ± 0.31	21.42 ± 0.22
	SR-FRP	**75.32** ± 0.35	2.33 ± 0.26	**75.74** ± 0.06	3.52 ± 0.24	**74.40** ± 0.03	8.75 ± 0.84
	SR-FedProx	70.49 ± 0.14	22.06 ± 0.08	72.91 ± 0.14	13.84 ± 0.05	71.64 ± 0.17	21.57 ± 0.39
	FedR^{2}DG	71.70 ± 0.19	**29.70** ± 0.23	72.68 ± 0.11	**26.18** ± 0.10	72.34 ± 0.20	**28.37** ± 0.19

Table 18: OfficeHome, target P: per-case accuracy (%), mean ± standard error.

		Defended A		Defended C		Defended R	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	57.76 ± 0.08	45.09 ± 0.27	55.89 ± 0.08	41.87 ± 0.48	72.20 ± 0.35	60.76 ± 0.38
	SR-FedAvg	74.37 ± 0.32	56.91 ± 0.31	74.75 ± 0.01	55.03 ± 0.17	75.35 ± 0.00	60.61 ± 0.17
	SR-FRP	**74.85** ± 0.29	53.12 ± 0.58	**75.56** ± 0.14	51.32 ± 0.26	**75.78** ± 0.07	58.54 ± 1.11
	SR-FedProx	74.55 ± 0.24	57.08 ± 0.34	74.89 ± 0.19	55.00 ± 0.09	75.29 ± 0.07	60.84 ± 0.30
	FedR^{2}DG	74.49 ± 0.08	**61.38** ± 0.08	74.04 ± 0.31	**61.20** ± 0.28	74.96 ± 0.34	**62.47** ± 0.48
4/255	Robust ERM	53.91 ± 0.28	34.51 ± 0.29	54.82 ± 0.41	33.71 ± 0.47	69.34 ± 0.34	52.24 ± 0.47
	SR-FedAvg	72.43 ± 0.12	43.54 ± 0.73	74.02 ± 0.10	38.07 ± 0.74	73.75 ± 0.16	50.72 ± 0.27
	SR-FRP	**73.70** ± 0.11	29.05 ± 0.58	**75.11** ± 0.32	26.32 ± 1.12	**75.23** ± 0.25	38.53 ± 1.22
	SR-FedProx	72.51 ± 0.15	44.00 ± 0.31	73.90 ± 0.14	38.08 ± 0.86	73.71 ± 0.13	50.81 ± 0.30
	FedR^{2}DG	72.77 ± 0.16	**51.69** ± 0.11	72.63 ± 0.40	**50.98** ± 0.02	73.39 ± 0.23	**53.58** ± 0.17
8/255	Robust ERM	46.00 ± 0.13	21.33 ± 0.37	51.75 ± 0.31	22.18 ± 0.10	64.41 ± 0.14	39.11 ± 0.35
	SR-FedAvg	68.87 ± 0.18	25.84 ± 0.79	72.01 ± 0.30	17.76 ± 0.29	70.38 ± 0.22	35.76 ± 0.62
	SR-FRP	**72.55** ± 0.01	4.82 ± 0.72	**74.31** ± 0.19	5.56 ± 0.39	**73.03** ± 0.22	12.94 ± 1.03
	SR-FedProx	68.25 ± 0.60	26.08 ± 0.39	72.14 ± 0.47	18.12 ± 0.51	70.51 ± 0.17	35.68 ± 0.51
	FedR^{2}DG	69.85 ± 0.32	**36.42** ± 0.24	70.36 ± 0.45	**34.24** ± 0.31	70.65 ± 0.48	**39.38** ± 0.29

D.4 MINI-DOMAINNET50

Domain codes: C = Clipart, P = Painting, R = Real, S = Sketch.

Table 19: mini-DomainNet50, target C: per-case accuracy (%), mean ± standard error.

		Defended P		Defended R		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	58.18 ± 0.04	47.71 ± 0.20	52.05 ± 0.45	41.21 ± 0.47	65.69 ± 0.32	54.26 ± 0.18
	SR-FedAvg	73.90 ± 0.31	59.63 ± 0.50	73.81 ± 0.37	57.49 ± 0.56	75.30 ± 0.29	56.65 ± 0.34
	SR-FRP	**74.90** ± 0.65	57.08 ± 0.54	**74.70** ± 0.20	56.92 ± 0.38	**76.03** ± 0.22	54.76 ± 0.37
	SR-FedProx	73.91 ± 0.33	59.50 ± 0.29	73.55 ± 0.15	57.27 ± 0.34	75.63 ± 0.22	57.11 ± 0.19
	FedR²DG	73.99 ± 0.58	**62.75** ± 0.44	73.95 ± 0.09	**62.16** ± 0.33	75.38 ± 0.25	**63.67** ± 0.39
4/255	Robust ERM	55.74 ± 0.13	40.51 ± 0.28	51.37 ± 0.43	36.32 ± 0.31	63.00 ± 0.17	46.59 ± 0.07
	SR-FedAvg	72.80 ± 0.37	50.93 ± 0.16	72.91 ± 0.25	48.65 ± 0.49	75.47 ± 0.05	45.10 ± 0.30
	SR-FRP	**74.84** ± 0.22	41.83 ± 1.32	**74.32** ± 0.48	39.12 ± 1.36	**75.89** ± 0.23	37.99 ± 0.99
	SR-FedProx	72.85 ± 0.32	51.35 ± 0.19	72.89 ± 0.51	48.27 ± 0.15	75.51 ± 0.26	44.88 ± 0.34
	FedR²DG	73.63 ± 0.47	**56.03** ± 0.40	73.32 ± 0.22	**55.39** ± 0.38	74.73 ± 0.20	**56.26** ± 0.38
8/255	Robust ERM	50.60 ± 1.48	30.54 ± 0.78	48.01 ± 0.47	28.55 ± 0.48	55.00 ± 2.32	33.52 ± 1.99
	SR-FedAvg	70.61 ± 0.24	39.09 ± 0.60	69.92 ± 0.23	34.53 ± 1.30	72.40 ± 0.37	29.10 ± 0.65
	SR-FRP	**73.65** ± 0.65	17.63 ± 0.24	**73.66** ± 0.37	16.69 ± 0.30	**74.95** ± 0.21	12.74 ± 1.41
	SR-FedProx	69.66 ± 1.54	38.07 ± 1.43	69.57 ± 0.85	32.73 ± 1.28	72.59 ± 0.31	29.53 ± 0.29
	FedR²DG	71.32 ± 0.42	**45.37** ± 0.42	71.43 ± 0.41	**44.82** ± 0.08	73.40 ± 0.18	**45.43** ± 0.21

Table 20: mini-DomainNet50, target P: per-case accuracy (%), mean ± standard error.

		Defended C		Defended R		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	47.67 ± 0.75	27.40 ± 0.49	65.05 ± 0.11	39.86 ± 0.40	49.05 ± 0.91	28.22 ± 0.30
	SR-FedAvg	70.98 ± 0.45	33.99 ± 0.24	72.17 ± 0.30	41.10 ± 0.21	70.65 ± 0.20	35.25 ± 0.36
	SR-FRP	**72.15** ± 0.13	26.75 ± 0.63	**72.49** ± 0.30	32.94 ± 1.00	**72.12** ± 0.24	30.24 ± 0.78
	SR-FedProx	70.93 ± 0.44	33.97 ± 0.27	72.12 ± 0.35	41.05 ± 0.32	70.59 ± 0.32	35.33 ± 0.61
	FedR²DG	70.52 ± 0.22	**44.19** ± 0.23	71.17 ± 0.34	**45.55** ± 0.07	69.85 ± 0.35	**44.67** ± 0.30
4/255	Robust ERM	44.15 ± 0.41	17.39 ± 0.29	59.71 ± 0.21	26.23 ± 0.37	42.31 ± 0.48	16.89 ± 0.13
	SR-FedAvg	68.79 ± 0.54	16.41 ± 0.23	69.33 ± 0.44	25.54 ± 0.25	65.99 ± 0.20	19.01 ± 0.30
	SR-FRP	**71.41** ± 0.19	7.83 ± 0.51	**71.63** ± 0.27	11.67 ± 0.40	**70.81** ± 0.33	10.77 ± 0.92
	SR-FedProx	68.94 ± 0.65	16.19 ± 0.21	69.53 ± 0.37	25.15 ± 0.45	66.09 ± 0.24	18.80 ± 0.50
	FedR²DG	67.66 ± 0.29	**27.63** ± 0.15	68.53 ± 0.21	**30.95** ± 0.18	67.16 ± 0.66	**28.88** ± 0.16
8/255	Robust ERM	38.25 ± 0.17	8.15 ± 0.13	51.18 ± 0.41	12.98 ± 0.24	32.27 ± 0.95	7.46 ± 0.51
	SR-FedAvg	63.04 ± 0.13	4.47 ± 0.12	63.25 ± 0.86	10.42 ± 0.27	57.90 ± 0.29	6.83 ± 0.19
	SR-FRP	**69.35** ± 0.67	0.75 ± 0.04	**69.28** ± 0.12	1.47 ± 0.18	**67.92** ± 0.19	1.47 ± 0.08
	SR-FedProx	62.71 ± 0.38	4.59 ± 0.28	62.61 ± 0.51	10.54 ± 0.16	57.95 ± 0.73	6.94 ± 0.16
	FedR²DG	63.13 ± 0.14	**11.84** ± 0.12	61.66 ± 0.16	**14.87** ± 0.10	62.31 ± 0.47	**13.01** ± 0.17

Table 21: mini-DomainNet50, target R: per-case accuracy (%), mean ± standard error.

		Defended C		Defended P		Defended S	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	59.43 ± 0.41	40.13 ± 0.39	72.41 ± 0.20	54.53 ± 0.24	59.21 ± 0.66	39.24 ± 0.67
	SR-FedAvg	79.14 ± 0.31	48.42 ± 0.25	78.95 ± 0.41	57.50 ± 0.45	78.04 ± 0.22	49.15 ± 0.05
	SR-FRP	**79.90** ± 0.09	41.25 ± 1.70	**79.36** ± 0.35	50.82 ± 0.48	**79.21** ± 0.37	47.10 ± 0.70
	SR-FedProx	79.29 ± 0.34	48.23 ± 0.43	78.74 ± 0.34	57.46 ± 0.26	77.98 ± 0.24	49.41 ± 0.10
	FedR^{2}DG	77.63 ± 0.25	**58.27** ± 0.32	78.27 ± 0.01	**60.32** ± 0.12	77.68 ± 0.22	**58.81** ± 0.44
4/255	Robust ERM	55.27 ± 0.40	27.97 ± 0.25	67.03 ± 0.21	39.55 ± 0.26	51.31 ± 0.43	24.87 ± 0.22
	SR-FedAvg	77.52 ± 0.23	26.89 ± 0.22	75.89 ± 0.33	40.51 ± 0.29	74.87 ± 0.23	28.82 ± 0.49
	SR-FRP	**78.43** ± 0.26	14.96 ± 0.27	**78.19** ± 0.27	22.40 ± 0.42	**78.13** ± 0.49	19.59 ± 0.82
	SR-FedProx	77.45 ± 0.06	26.91 ± 0.12	76.11 ± 0.18	40.68 ± 0.32	74.69 ± 0.34	28.68 ± 0.66
	FedR^{2}DG	75.69 ± 0.17	**41.29** ± 0.17	75.72 ± 0.18	**45.36** ± 0.39	75.54 ± 0.32	**42.45** ± 0.29
8/255	Robust ERM	48.22 ± 0.34	15.67 ± 0.03	59.01 ± 0.29	21.83 ± 0.17	39.17 ± 1.04	12.89 ± 0.36
	SR-FedAvg	72.52 ± 0.59	8.30 ± 0.11	69.61 ± 1.10	19.09 ± 1.20	68.46 ± 0.31	11.61 ± 0.23
	SR-FRP	**77.15** ± 0.25	2.01 ± 0.08	**76.75** ± 0.19	4.31 ± 0.48	**75.95** ± 0.29	3.11 ± 0.18
	SR-FedProx	72.11 ± 0.53	9.03 ± 0.64	70.11 ± 1.19	19.26 ± 0.98	68.40 ± 0.28	11.77 ± 0.27
	FedR^{2}DG	72.81 ± 0.16	**20.02** ± 0.09	70.00 ± 0.28	**24.47** ± 0.18	71.13 ± 0.07	**20.93** ± 0.23

Table 22: mini-DomainNet50, target S: per-case accuracy (%), mean ± standard error.

		Defended C		Defended P		Defended R	
ε	Method	Std.	Rob.	Std.	Rob.	Std.	Rob.
2/255	Robust ERM	58.67 ± 0.22	44.63 ± 0.36	55.97 ± 0.65	41.76 ± 0.73	48.45 ± 0.60	33.50 ± 0.57
	SR-FedAvg	72.35 ± 0.33	47.11 ± 0.49	70.45 ± 0.38	52.79 ± 0.31	70.54 ± 0.61	49.34 ± 0.48
	SR-FRP	72.05 ± 0.27	42.96 ± 1.03	**71.17** ± 0.20	50.27 ± 0.63	**71.24** ± 0.28	47.20 ± 1.52
	SR-FedProx	**72.37** ± 0.46	47.19 ± 0.38	70.45 ± 0.45	52.69 ± 0.35	70.61 ± 0.73	49.29 ± 0.29
	FedR^{2}DG	70.87 ± 0.42	**54.31** ± 0.25	70.96 ± 0.25	**54.57** ± 0.24	70.49 ± 0.09	**53.56** ± 0.08
4/255	Robust ERM	56.74 ± 0.26	36.86 ± 0.08	51.13 ± 0.47	31.39 ± 0.41	45.77 ± 0.44	24.83 ± 0.18
	SR-FedAvg	71.65 ± 0.24	30.71 ± 0.60	67.45 ± 0.24	40.35 ± 0.13	67.89 ± 0.42	36.75 ± 0.34
	SR-FRP	**71.67** ± 0.42	23.37 ± 0.69	**69.92** ± 0.28	32.36 ± 0.68	**70.26** ± 0.17	28.90 ± 1.43
	SR-FedProx	71.49 ± 0.26	30.87 ± 0.65	67.17 ± 0.91	40.21 ± 0.46	67.99 ± 0.43	36.39 ± 0.28
	FedR^{2}DG	69.46 ± 0.30	**43.43** ± 0.23	68.75 ± 0.13	**44.01** ± 0.34	68.80 ± 0.50	**42.55** ± 0.30
8/255	Robust ERM	52.59 ± 0.54	26.57 ± 0.17	44.17 ± 0.70	19.77 ± 0.36	39.36 ± 0.11	15.47 ± 0.10
	SR-FedAvg	70.15 ± 0.24	15.01 ± 0.51	62.81 ± 1.43	26.39 ± 0.27	63.68 ± 0.39	21.75 ± 0.72
	SR-FRP	**71.31** ± 0.16	6.29 ± 0.81	**67.88** ± 0.41	13.46 ± 0.38	**68.15** ± 0.06	10.39 ± 0.73
	SR-FedProx	69.81 ± 0.21	15.67 ± 1.02	62.78 ± 1.32	26.03 ± 0.83	63.52 ± 0.37	21.39 ± 0.22
	FedR^{2}DG	67.34 ± 0.27	**28.69** ± 0.17	64.23 ± 0.14	**30.18** ± 0.16	64.12 ± 0.49	**28.41** ± 0.30

E Reproducibility Details

We document the experimental environment, random seeds, training configuration, data processing, and evaluation protocol used in the experiments.

E.1 Experimental Environment and Randomness

Each training run was conducted on a single NVIDIA A100 80GB PCIe GPU. The experiments were run on a local server equipped with an AMD EPYC 7413 24-Core Processor and approximately 1.26 TB of system memory. The operating system was Ubuntu 20.04.3 LTS. Our implementation used Python 3.9.7, PyTorch 1.11.0, torchvision 0.12.0, CUDA 11.3, and cuDNN 8.2.0.

All reported main results were obtained from three independent runs using random seeds $\{1, 42, 123\}$. The random seeds controlled data-loader shuffling, mini-batch sampling, stochastic data augmentation, PGD random initialization, and other stochastic training operations. All compared methods used the same set of random seeds for each experimental case. Minor nondeterminism may remain due to GPU kernels.

For each seed, we first average the results over the 12 experimental cases formed by four held-out target domains and three defended-source choices. We then report the mean and standard error over the three seed-level averages. For seed-level results $\{a_s\}_{s=1}^{3}$, the reported standard error is

$$\mathrm{SE} = \frac{\mathrm{Std}(a_1, a_2, a_3)}{\sqrt{3}}.$$

E.2 Complete Experimental Configuration

Unless otherwise stated, the same configuration is used across all datasets, held-out target domains, defended-source choices, perturbation budgets, and random seeds.

Table 23: Hyperparameters used in the main experiments.

Hyperparameter	Value
Backbone	ResNet-50
Initialization	Robust ImageNet model pretrained at $\varepsilon = 1/255$
Optimizer	SGD
Learning rate	0.001
Momentum	0.9
Weight decay	5×10^{-4}
Batch size	32
Communication rounds	250
Warm-up rounds	50
Main-training rounds	200
Local steps per round	20
Total local steps per client	5,000
PGD steps	10
Training perturbation budgets	$\varepsilon \in \{2, 4, 8\}/255$
PGD step size	$\varepsilon/4$
PGD random start	Enabled
Clean steps on weakly defended clients	$B_C = 5$
Replay mini-batches	$B_H = 5$
Replays per mini-batch	$\zeta = 3$
Step-level EMA coefficient	$\gamma_{\mathrm{ema}} = 0.1$
Prototype loss weight	$\lambda = 1.0$
Prototype temperature	$\tau_{\mathrm{con}} = 0.5$
Centroid EMA coefficient	$\alpha_{\mathrm{ema}} = 0.9$
Centroid update interval	$I_{\mathrm{cent}} = 5$
FedProx coefficient	$\mu = 0.1$
Adapted FRP coefficient	$\lambda_{\mathrm{FRP}} = 0.5$
Robust evaluation	AA
Evaluation perturbation budgets	$\varepsilon \in \{2, 4, 8\}/255$
Random seeds	$\{1, 42, 123\}$

E.3 Data Processing and Evaluation

E.3.1 Evaluation Protocol.

Standard accuracy is the top-1 classification accuracy on clean held-out target images, and robust accuracy is the top-1 classification accuracy under AA at the specified ℓ_∞ perturbation budget. We report their harmonic mean as

$$H = \frac{2A_{\text{std}}A_{\text{rob}}}{A_{\text{std}} + A_{\text{rob}}},$$

which summarizes the standard–robust balance and penalizes methods that perform well on only one of the two criteria.

E.3.2 mini-DomainNet50 Construction.

We construct mini-DomainNet50 from the original-resolution DomainNet images (Peng et al., 2019) using the Mini DomainNet image lists (Zhou et al., 2021). We consider four domains: Clipart, Painting, Real, and Sketch.

The class-selection stage uses all four domains to define a fixed benchmark. It precedes the LODO training/evaluation protocol and is not a target-dependent training operation. Candidate classes are enumerated from the Mini DomainNet Clipart training split. For each candidate class, we count the available original-resolution DomainNet images in all four domains. Classes containing fewer than 20 images in any domain are excluded. Among the remaining classes, we select the 50 classes with the largest total image counts across the four domains.

For the selected classes, we combine the Mini DomainNet training and test lists within each domain and sample 5,000 images without replacement using class-proportional stratified sampling. The complete class list and the resulting class-wise image counts are reported in Table 24 and Figure 3, respectively.

Table 24: The 50 classes included in mini-DomainNet50.

Index	Class	Index	Class
0	The_Eiffel_Tower	25	onion
1	asparagus	26	penguin
2	bear	27	raccoon
3	bird	28	saxophone
4	bottlecap	29	sea_turtle
5	butterfly	30	see_saw
6	canoe	31	sheep
7	cat	32	shoe
8	dog	33	skateboard
9	drums	34	snake
10	duck	35	speedboat
11	dumbbell	36	spider
12	elephant	37	squirrel
13	feather	38	strawberry
14	fish	39	streetlight
15	flower	40	submarine
16	frog	41	swan
17	grapes	42	table
18	helicopter	43	teapot
19	horse	44	tiger
20	lion	45	train
21	lobster	46	vase
22	monkey	47	watermelon
23	mug	48	whale
24	mushroom	49	zebra

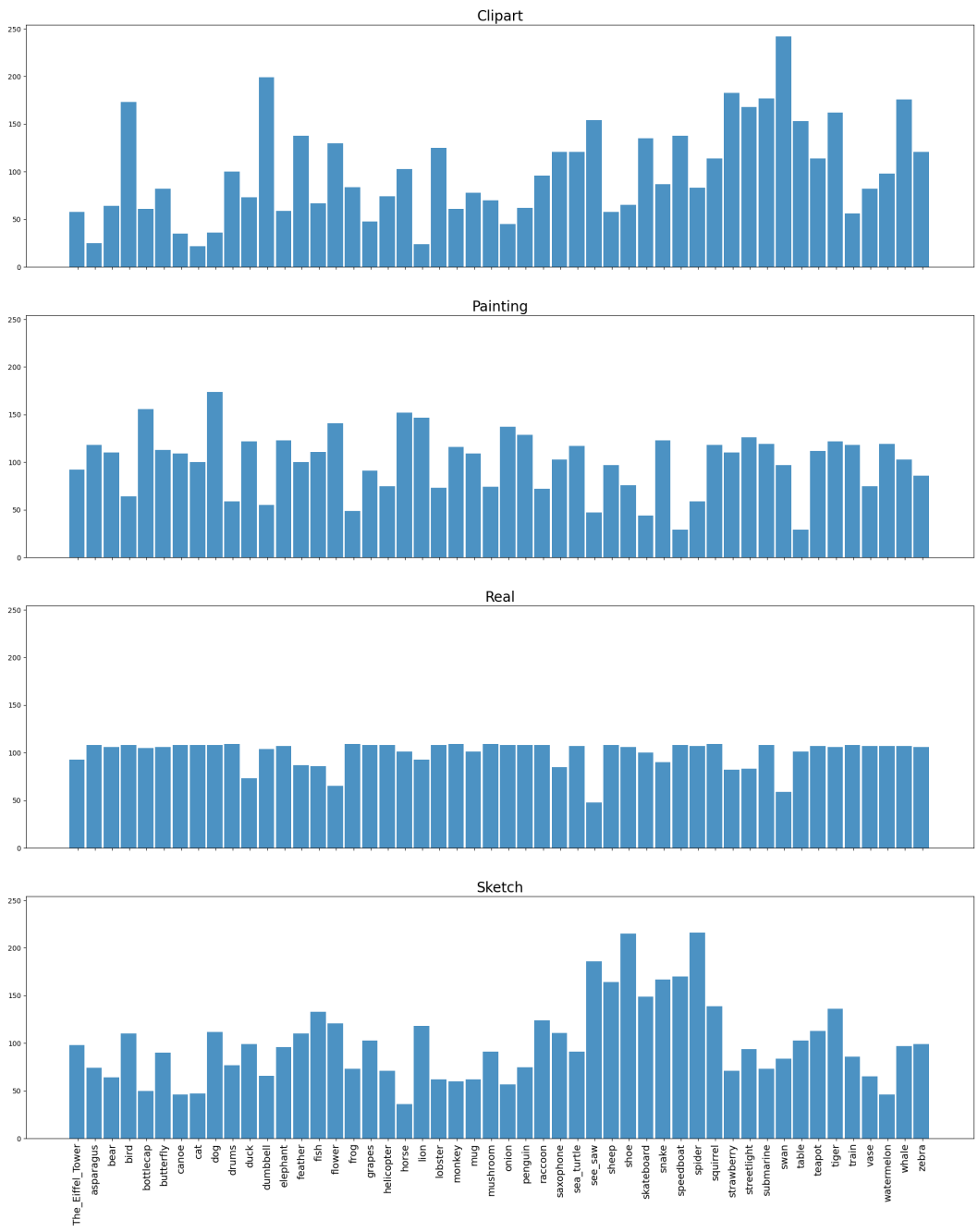


Figure 3: Class-wise image counts in the final mini-DomainNet50 dataset across Clipart, Painting, Real, and Sketch, with 5,000 images per domain.

E.4 Fixed Hyperparameter Configuration

To avoid target-dependent or case-specific tuning in the domain generalization setting, we use a single fixed configuration across all datasets, held-out domains, defended-source choices, and perturbation budgets. We set the prototype loss weight to $\lambda = 1.0$, corresponding to an unscaled auxiliary-loss coefficient, and use the fixed contrastive temperature $\tau_{\text{con}} = 0.5$. We set the centroid update interval to $I_{\text{cent}} = 5$ based on computational considerations, reducing the frequency of centroid recomputation while retaining regular prototype updates.

In the post-hoc sensitivity analysis of Section 5.3, we evaluate the prototype loss weight over

$$\lambda \in \{0, 0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 1.7, 2.0\},$$

the prototype temperature over

$$\tau_{\text{con}} \in \{0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 1.7, 2.0\},$$

and the centroid update interval over

$$I_{\text{cent}} \in \{1, 3, 5, 10, 20\}.$$

The experiments in Section 5.3 were not used to select the configuration of the main experiments.