FEDNASC: NOISE-AWARE SEMANTIC CONTROL FOR FEDERATED LORA ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated LoRA adaptation of pretrained Vision Transformers must distinguish genuine cross-client semantic variation from finite-sample prototype noise. This is difficult when local class support is small, because unstable prototypes can mimic genuine within-class domain modes. We propose FEDNASC, a noise-aware semantic controller that explicitly estimates this instability from split-half prototype disagreement. Each client augments its class prototype with a split-half instability statistic, and the server forms a matched class-wise noise floor, subtracts it from observed dispersion, and maps the calibrated excess to a score controlling prototype-bank capacity, classifier anchoring, semantic regularization, and prototype-readout fusion. Under a fixed-map perturbation model, we show that the untruncated calibrated difference estimates population representation dispersion in expectation up to controlled noise-dependent terms. Population dispersion in turn determines the weighted Euclidean single-prototype consensus error and bounds its unit-norm counterpart. We further derive a stationarity bound for the implemented optimization and server pipeline with explicit implementation residuals. Across CIFAR-100, HAM10000, DomainNet, and NICO++, FEDNASC outperforms the strongest of six federated adaptation baselines by 1.4–2.4 percentage points across all ten settings. Ablations show that noise calibration contributes most under label skew, whereas adaptive bank capacity contributes most under domain shift, consistent with the intended roles of the two mechanisms.

1 INTRODUCTION

Pretrained Vision Transformers transfer effectively across visual recognition tasks (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021b; He et al., 2022), but their adaptation data are often distributed across hospitals, institutions, or devices that cannot pool raw images (McMahan et al., 2017; Kairouz & McMahan, 2021). Federated adaptation therefore increasingly relies on LoRA and related parameter-efficient methods (Houlsby et al., 2019; Jia et al., 2022; Hu et al., 2021). Existing federated LoRA methods primarily study how low-rank updates are initialized, shared, aggregated, or personalized (Zhang et al., 2024c; Sun et al., 2024; Wang et al., 2024b; Guo et al., 2025; Bian et al., 2025; Peng et al., 2026), rather than whether class-level semantic variation observed across clients is reliable enough to preserve.

Cross-client prototype dispersion can arise from two fundamentally different sources. Under domain shift, a class may genuinely occupy several regions of feature space, so collapsing them into a single consensus direction discards useful semantic structure. Under label skew, however, a client with few examples may produce an unstable prototype even when the underlying semantics are shared, so treating that disagreement as semantic evidence promotes sampling noise. The two cases demand opposite treatment yet both appear as elevated prototype dispersion.

Local support alone cannot resolve this ambiguity because prototype instability depends jointly on sample count and within-class feature variability. A fixed support threshold may therefore reject reliable low-variance prototypes while retaining noisy high-variance ones. In our controlled zero-separation diagnostic (Section 5.3), raw dispersion is 0.1130 at support five but only 0.0063 at support one hundred, despite the true population dispersion being zero. Raw dispersion can therefore substantially overstate semantic heterogeneity in low-support regimes.

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.

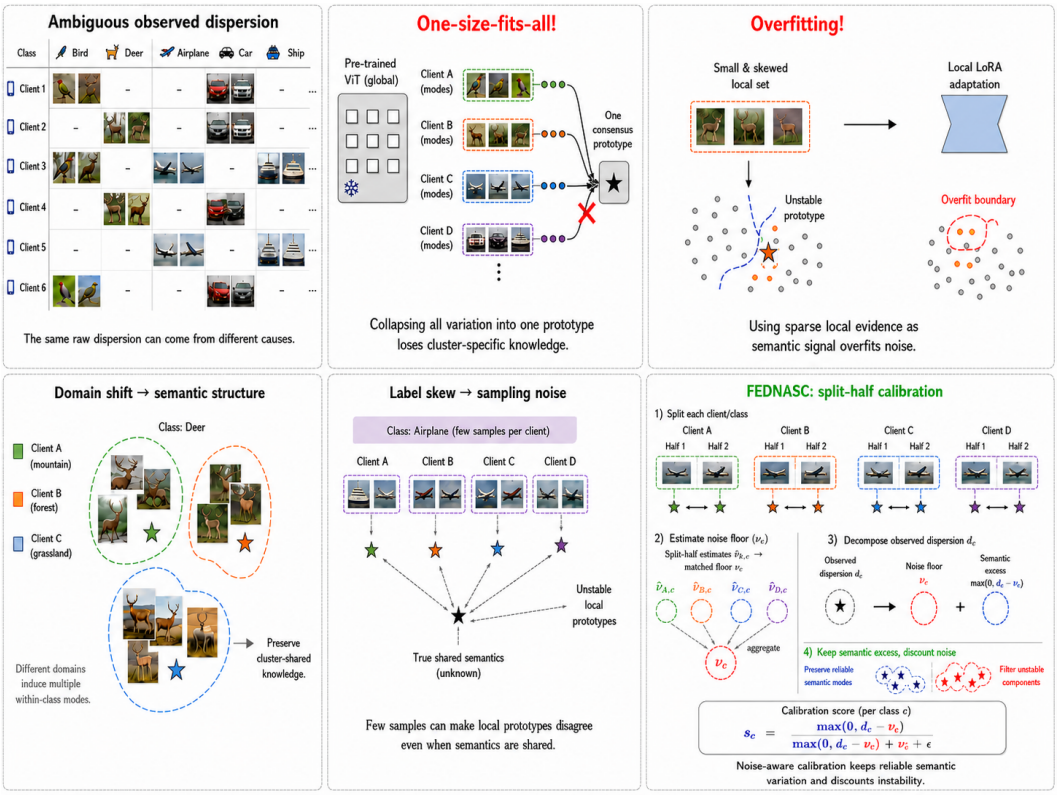


Figure 1: The same raw cross-client prototype dispersion can reflect either genuine semantic structure or finite-sample noise. Under domain shift, collapsing distinct class modes into one prototype loses useful structure; under label skew, sparse local observations can create spurious disagreement. FEDNASC estimates a matched noise floor ν_c from split-half disagreement, subtracts it from the observed dispersion d_c, and retains the calibrated excess through the score s_c.

Prototype-based federated methods communicate class-level information efficiently and have been extended with trainable global prototypes, normalized classifier heads, variance-aware aggregation, and richer distributional representations (Tan et al., 2022a;b; Zhang et al., 2024b; Dai et al., 2023; Wang et al., 2024a; Zhang et al., 2024a; Jin et al., 2025; Ma et al., 2025; Lu et al., 2026). These approaches model prototype structure, whereas FEDNASC explicitly separates a matched finite-sample noise component from observed cross-client dispersion before using the residual variation for semantic control. Figure 1 illustrates this distinction.

We propose FEDNASC, a class-wise semantic controller for federated LoRA adaptation that calibrates cross-client prototype dispersion before using it as semantic evidence. Each client uploads a class prototype together with a split-half statistic measuring its finite-sample instability. The server estimates a matched noise floor ν_c, subtracts it from the observed dispersion d_c, and maps the calibrated excess to a score s_c. This score jointly regulates prototype-bank capacity, classifier anchoring, client-side semantic guidance, and prototype-readout fusion, giving classes with reliable semantic variation greater flexibility while treating noise-dominated disagreement more conservatively.

Our contributions are summarized as follows.

- We introduce a split-half calibration procedure that estimates the finite-sample instability of class prototypes from client-side statistics. By subtracting the matched noise floor from cross-client prototype dispersion, FEDNASC derives a class-wise score that reflects calibrated semantic variation rather than raw disagreement.
- We develop a semantic controller in which the calibrated score governs adaptive bank capacity, classifier anchoring, semantic guidance, and prototype-readout fusion. The analysis

links calibrated dispersion to population dispersion, multi-prototype representation error, and score-dependent classifier displacement.

- We evaluate FEDNASC for federated ViT-LoRA adaptation under label skew, global long-tail imbalance, domain shift, and context shift. Estimator diagnostics validate the proposed calibration, while functional ablations identify the distinct contributions of noise calibration and adaptive semantic control.

2 RELATED WORK

Federated parameter-efficient adaptation. Statistical heterogeneity has motivated federated optimization methods based on proximal regularization, control variates, normalized aggregation, and client-local normalization (Li et al., 2020; Karimireddy et al., 2020; Wang et al., 2020; Acar et al., 2021; Li et al., 2021c). These challenges become particularly costly when adapting large pretrained backbones, motivating parameter-efficient alternatives that update only lightweight modules such as adapters, visual prompts, feature-scaling parameters, or low-rank factors (Houlsby et al., 2019; Jia et al., 2022; Chen et al., 2022; Lian et al., 2022; Hu et al., 2021). Federated LoRA methods further address how such low-rank updates should be trained and aggregated across clients. Some target factor-wise aggregation inconsistency by freezing one factor (Sun et al., 2024), sharing only one factor across clients (Guo et al., 2025), or refining initialization and aggregation (Singhal et al., 2024; Bian et al., 2025). Others accommodate variation in rank, task, or resource budget (Wang et al., 2024b; Bai et al., 2024), or introduce personalization and latent client structure (Zhang et al., 2024c; Bian et al., 2026; Fan et al., 2025; Peng et al., 2026). SpecTraL (Ramesh & Dass, 2026) further separates consensus signal from non-IID noise in the LoRA spectral space for layer-wise rank discovery. These methods operate primarily on the parameter pathway, whereas FEDNASC calibrates the reliability of class-level semantic variation.

Prototypes and the reliability of semantic evidence. Prototypes provide federated learning with a compact channel for class-level semantics. FedProto exchanges class-wise prototypes to align heterogeneous local objectives (Tan et al., 2022a), while subsequent work makes global prototypes trainable (Zhang et al., 2024b), couples prototype transfer with contrastive objectives (Tan et al., 2022b), regularizes classifier geometry (Dai et al., 2023), models cross-domain representation variance (Wang et al., 2024a; Zhang et al., 2024a), and extends prototype exchange to distillation, conservation, and geometry-guided alignment (Jin et al., 2025; Zhang et al., 2025; Ma et al., 2025). GFPL (Lu et al., 2026) uses Gaussian mixture models to represent class-wise feature distributions under data imbalance. While these approaches capture variation in class representations, finite-sample instability can also contribute to cross-client prototype disagreement, particularly under sparse local support. Related personalization, representation-sharing, and federated domain-generalization methods address heterogeneity through local specialization or shared representations (Arivazhagan et al., 2019; Li et al., 2021a;b; Collins et al., 2021; Liu et al., 2021a; Li et al., 2025). In contrast, FEDNASC estimates a matched noise floor from client-side split-half statistics and uses the calibrated residual dispersion to control downstream semantic operations.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We consider federated classification over K clients with label space $\mathcal{Y} = \{1, \dots, C\}$. Client k holds a local dataset $\mathcal{D}_k \sim \mathcal{P}_k$, with $n_k = |\mathcal{D}_k|$ samples and $n_{k,c}$ samples from class c. Local adaptation and semantic statistics use the same partition.

A pretrained ViT is adapted through LoRA, the classifier, and selected LayerNorm affine parameters. The trainable client state is

$$\omega_k = (\Delta_k, \phi_k), \qquad \phi_k = (W_k, b_k, \psi_k),$$

and the adapted encoder produces

$$\widetilde{f}_k(x) = \mathrm{Norm}(f_k(x)), \quad z_{\mathrm{lin},k}(x) = W_k f_k(x) + b_k, \quad q_{\mathrm{lin},k}(x) = \mathrm{softmax}(z_{\mathrm{lin},k}(x)).$$

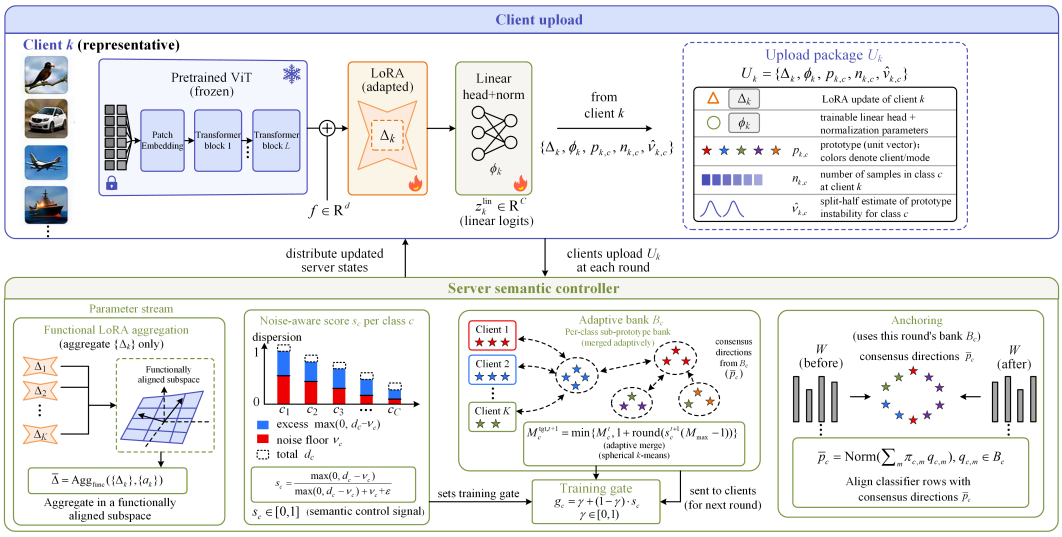


Figure 2: Overview of the FEDNASC framework. Clients perform local LoRA adaptation with score-controlled semantic guidance and upload model updates together with class prototypes, support counts, and split-half instability estimates. The server calibrates cross-client prototype dispersion against the estimated noise floor to obtain a class-wise score s_c, which controls prototype-bank capacity, classifier anchoring, and the semantic gate. After the final round, a common-reference refresh constructs the final semantic state for inference.

Normalization and degenerate cases are defined in Eq. (19) and Appendix A.

Figure 2 summarizes the framework. At round t, the server samples clients $\mathcal{S}^t$ and broadcasts $\{\bar{\Delta}^t, \bar{\phi}^t, \mathcal{B}^t, \mathbf{g}^t\}$. Clients use the broadcast model and semantic state during local adaptation and return class prototypes with split-half instability estimates. The server calibrates cross-client prototype dispersion and updates the semantic state for the next round.

3.2 CLIENT-SIDE ADAPTATION AND SEMANTIC ESTIMATION

Each selected client initializes from $(\bar{\Delta}^t, \bar{\phi}^t)$ and uses the fixed prototype bank $\mathcal{B}^t$ for semantic guidance. Its strength is controlled by the broadcast gate g_c^t and local class support:

$$\begin{aligned} \delta_{k,c} &= \exp(-n_{k,c}/n_0)\mathbf{1}[n_{k,c} > 0], \\ g_{k,c}^{\mathrm{eff},t} &= g_c^t\Big[(1-\rho_t) + \rho_t\big\{\kappa + (1-\kappa)\delta_{k,c}\big\}\Big], \end{aligned} \tag{1}$$

with scarcity scale n_0, warm-up ramp ρ_t, and floor κ.

Let $\mathcal{C}_{\mathcal{B}}^t = \{c : \mathcal{B}_c^t \neq \emptyset\}$. For each covered class,

$$u_{k,c}^t(x) = -\log|\mathcal{B}_c^t| + \log \sum_{(q,\pi)\in\mathcal{B}_c^t} \exp\left(\frac{\langle \widetilde{f}_k^t(x), q\rangle}{\tau}\right). \tag{2}$$

The log-mean-exp removes the advantage of larger banks, while omitting stored support π gives retained modes equal within-class status. For covered targets,

$$\ell_{\mathrm{proto}}^t(x,y) = -u_{k,y}^t(x) + \log \sum_{c\in\mathcal{C}_{\mathcal{B}}^t} \exp(u_{k,c}^t(x)),$$

and

$$\mathcal{L}_k^{\mathrm{sem},t} = \frac{1}{N_k^t} \sum_{(x,y)\in\mathcal{M}_k} \mathbf{1}[y \in \mathcal{C}_{\mathcal{B}}^t] g_{k,y}^{\mathrm{eff},t} \ell_{\mathrm{proto}}^t(x,y), \tag{3}$$

where $\mathcal{M}_k$ denotes the current local minibatch and

$$N_k^t = \sum_{(x,y)\in\mathcal{M}_k} \mathbf{1}[y \in \mathcal{C}_{\mathcal{B}}^t] + \epsilon.$$

After local optimization, client k forms a prototype for each class with $n_{k,c} \geq n_{\min}$,

$$p_{k,c}^t = \mathrm{Norm}\left(\frac{1}{n_{k,c}} \sum_{(x,y)\in\mathcal{D}_k:\, y=c} \widetilde{f}_k^t(x)\right). \tag{4}$$

To estimate its finite-sample instability, the class examples are split uniformly at random into two near-balanced subsets of sizes $m_{k,c}^{(a)} = \lfloor n_{k,c}/2 \rfloor$ and $m_{k,c}^{(b)} = \lceil n_{k,c}/2 \rceil$. Their normalized half-prototypes give

$$\widehat{\nu}_{k,c}^t = \chi_{k,c}\left(1 - \langle p_{k,c}^{(a)}, p_{k,c}^{(b)} \rangle\right), \qquad \chi_{k,c} = \frac{2m_{k,c}^{(a)} m_{k,c}^{(b)}}{n_{k,c}^2}. \tag{5}$$

For equal halves, $\chi_{k,c} = 1/2$. We set $n_{\min} = 5$. Under the fixed-map model of Section 4,

$$\mathbb{E}[\widehat{\nu}_{k,c}] = 2\nu_{k,c} + O(\nu_{k,c}^{3/2}),$$

where the round index is suppressed for the population noise. The analysis assumes conditionally i.i.d. within-class observations; grouped data should be split at the independent-unit level. The client uploads Δ_k^{t+1}, ϕ_k^{t+1}, and $(p_{k,c}^t, n_{k,c}, \widehat{\nu}_{k,c}^t)$ for each valid class, without a trainable prototype head or client-side multi-prototype module.

3.3 Server-Side Noise Calibration and Semantic Control

With $a_k^t = n_k / \sum_{j\in\mathcal{S}^t} n_j$, the server forms $\widetilde{W}^{t+1} = \sum_{k\in\mathcal{S}^t} a_k^t W_k^{t+1}$ and analogously averages the bias and adapted LayerNorm parameters. LoRA states are aggregated in function space rather than factor by factor. Writing $\mathrm{Dense}(\Delta_k) = B_k A_k$,

$$\Delta W_{\mathrm{avg}}^{t+1} = \sum_{k\in\mathcal{S}^t} a_k^t \,\mathrm{Dense}(\Delta_k^{t+1}), \qquad \bar{A}^{t+1} = \sum_{k\in\mathcal{S}^t} a_k^t A_k^{t+1}, \tag{6}$$

and

$$\bar{B}^{t+1} = \arg\min_B \left\{ \|B\bar{A}^{t+1} - \Delta W_{\mathrm{avg}}^{t+1}\|_F^2 + \lambda_{\mathrm{r}} \|B\|_F^2 \right\}. \tag{7}$$

The pair is rebalanced as specified in Eq. (20). Because this constrained reconstruction need not attain the best rank-r approximation, its residual is recorded explicitly. The semantic pathway itself is independent of the particular LoRA aggregation rule.

In parallel, the returned semantic statistics are used for noise calibration. Let $\mathcal{R}_c^t = \{p_{k,c}^t : k \in \mathcal{S}^t, n_{k,c} \geq n_{\min}\}$ and $M_c^t = |\mathcal{R}_c^t|$. We use square-root support weights

$$w_{k,c}^t = \frac{\sqrt{n_{k,c}}}{\sum_{j:p_{j,c}^t\in\mathcal{R}_c^t} \sqrt{n_{j,c}}}, \qquad Z_c^t = 1 - \sum_i (w_{i,c}^t)^2. \tag{8}$$

For $M_c^t \geq 2$, the instantaneous dispersion and matched noise floor are

$$\widetilde{d}_c^t = \frac{1}{\max\{Z_c^t, \epsilon\}} \sum_{i\neq j} w_{i,c}^t w_{j,c}^t \left(1 - \langle p_{i,c}^t, p_{j,c}^t \rangle\right), \tag{9}$$

$$\widetilde{\nu}_c^t = \frac{\sum_i w_{i,c}^t (1 - w_{i,c}^t) \widehat{\nu}_{i,c}^t}{\max\{Z_c^t, \epsilon\}}. \tag{10}$$

The common normalization places both quantities on the same scale. Their exponential moving averages are

$$(d_c^{t+1}, \nu_c^{t+1}) = \mu_{\mathrm{ema}}(d_c^t, \nu_c^t) + (1 - \mu_{\mathrm{ema}})(\widetilde{d}_c^t, \widetilde{\nu}_c^t), \tag{11}$$

updated when at least two valid prototypes arrive and otherwise retained. The calibrated excess and score are

$$e_c^{t+1} = [d_c^{t+1} - \nu_c^{t+1}]_+, \qquad s_c^{t+1} = \frac{e_c^{t+1}}{e_c^{t+1} + \nu_c^{t+1} + \epsilon}. \tag{12}$$

High scores indicate dispersion that survives noise calibration and low scores indicate disagreement largely explained by the estimated noise floor.

Score-driven control. The score sets a desired bank capacity, capped by the number of valid current prototypes,

$$\widetilde{M}_c^{t+1} = 1 + \left\lfloor s_c^{t+1}(M_{\max} - 1) + \tfrac{1}{2} \right\rfloor, \qquad M_c^{\mathrm{tgt},t+1} = \min\{M_c^t, \widetilde{M}_c^{t+1}\}. \tag{13}$$

Under linear support weights $r_{k,c}^t = n_{k,c} / \sum_{j:p_{j,c}^t \in \mathcal{R}_c^t} n_{j,c}$, the target spherical clustering objective is

$$\mathcal{Q}_c^{t+1} = \arg \max_{\substack{\mathcal{Q} \subset \mathbb{S}^{d-1} \\ |\mathcal{Q}| = M_c^{\mathrm{tgt},t+1}}} \sum_{k:p_{k,c}^t \in \mathcal{R}_c^t} r_{k,c}^t \max_{q \in \mathcal{Q}} \langle p_{k,c}^t, q \rangle. \tag{14}$$

The implementation approximates this objective with weighted spherical k-means (Appendix B.1). Each representative is stored with its assigned sample support; classes without a valid current prototype retain their previous bank.

The same score controls classifier anchoring toward

$$\bar{p}_c^{t+1} = \mathrm{Norm}\left(\sum_m \pi_{c,m}^{t+1} q_{c,m}^{t+1} \right),$$

where $\pi_{c,m}^{t+1}$ is the sample support assigned to representative $q_{c,m}^{t+1}$:

$$\begin{aligned} \eta_c^{t+1} &= \eta_0 (1 - s_c^{t+1}), \\ \breve{w}_c^{t+1} &= (1 - \eta_c^{t+1}) \, \mathrm{Norm}(\widetilde{W}_c^{t+1}) + \eta_c^{t+1} \bar{p}_c^{t+1}, \\ \bar{W}_c^{t+1} &= \|\widetilde{W}_c^{t+1}\|_2 \, \mathrm{Norm}(\breve{w}_c^{t+1}). \end{aligned} \tag{15}$$

Low-score classes receive stronger consensus anchoring, whereas high-score classes retain greater multi-representative flexibility. The next-round gate is

$$g_c^{t+1} = \gamma + (1 - \gamma) s_c^{t+1}, \qquad \gamma \in [0, 1),$$

and the server broadcasts $\{\bar{\Delta}^{t+1}, \bar{\phi}^{t+1}, \mathcal{B}^{t+1}, \mathbf{g}^{t+1}\}$.

Final semantic state and inference. After the final aggregation, all clients perform one no-gradient refresh under the frozen reference encoder

$$\widetilde{f}_{\mathrm{ref}}^{\star}(x) = \mathrm{Norm}(F_{\theta_0, \bar{\Delta}^T, \bar{\psi}^T}(x)).$$

The server discards the training-round banks and recomputes the final scores $s_c^\star$ and bank $\mathcal{B}^\star$ from the refreshed statistics, ensuring that inference prototypes share the reference representation space. The corresponding gate is $g_c^\star = \gamma + (1 - \gamma) s_c^\star$, after which each client performs five local fine-tuning epochs from $(\bar{\Delta}^T, \bar{\phi}^T)$ with $(\mathcal{B}^\star, \mathbf{g}^\star)$ fixed. This stage provides local personalization rather than additional federated optimization.

Each client then combines its linear readout $q_{\mathrm{lin},k}$ with the coverage-safe prototype readout $q_{\mathrm{proto},k}$ of Eq. (26). The score sets a prior for the mixing weight,

$$q_k^{(\alpha)} = \alpha q_{\mathrm{lin},k} + (1 - \alpha) q_{\mathrm{proto},k}, \qquad \widehat{\alpha}_k = \mathrm{clamp}(1 - \beta \bar{s}_k, \alpha_{\min}, 1), \tag{16}$$

where $\bar{s}_k$ is the support-weighted mean of $s_c^\star$ over observed classes. We select

$$\mathcal{A}_k = \{\mathrm{clamp}(\widehat{\alpha}_k - 0.1, \alpha_{\min}, 1), \widehat{\alpha}_k, \mathrm{clamp}(\widehat{\alpha}_k + 0.1, \alpha_{\min}, 1), 1\}, \qquad \alpha_k^\star = \max \arg \max_{\alpha \in \mathcal{A}_k} \mathrm{Acc}_{\mathcal{D}_k}(q_k^{(\alpha)}). \tag{17}$$

No test data are used for selection; an empty final bank sets $\alpha_k^\star = 1$. Appendix A gives the complete procedure and Appendix B.1 the shared constants.

3.4 Overall Training Objective and Closed-Loop Update

Combining the standard classification objective with the semantic term in Eq. (3), each selected client minimizes

$$\mathcal{L}_k^t = \mathcal{L}_k^{\mathrm{ce},t} + \lambda \mathcal{L}_k^{\mathrm{sem},t}. \tag{18}$$

Here, $\mathcal{L}_k^{\mathrm{ce},t}$ preserves the standard local classification objective, while $\mathcal{L}_k^{\mathrm{sem},t}$ introduces class-wise semantic guidance from the server. Its strength is controlled by $g_{k,c}^{\mathrm{eff},t}$, which combines the score-derived broadcast gate with local class support.

After local optimization, the returned prototypes and split-half statistics update d_c^{t+1}, ν_c^{t+1}, and hence s_c^{t+1}. The new score then determines the prototype-bank capacity, classifier anchoring, and next-round semantic gate. Thus, local adaptation and server-side calibration form the closed semantic-control loop summarized in Figure 2.

4 Theoretical Analysis

We state the main results underlying noise calibration, multi-prototype representation, and score-driven control. Let $\mu_{k,c}$ denote client k's population direction for class c, $\nu_{k,c} = \mathbb{E}[1 - \langle p_{k,c}, \mu_{k,c}\rangle]$ its prototype noise, $\nu_{c,\max} = \max_{k\in\mathcal{K}_c^t} \nu_{k,c}$, and $D_c^\star$ the weighted population angular dispersion. Complete assumptions and proofs are given in Appendix C.

Theorem 1 (Noise-calibrated dispersion)**.** *Under the fixed-map perturbation model,*

$$\mathbb{E}[\widehat{\nu}_{k,c}] = 2\nu_{k,c} + O(\nu_{k,c}^{3/2}),$$

and, for $M_c > 1$,

$$\mathbb{E}[\widetilde{d}_c - \widetilde{\nu}_c] = D_c^\star + O(\nu_{c,\max} D_c^\star) + O(\nu_{c,\max}^{3/2}).$$

The fitted-and-reused encoder case is given in Appendix C.

Theorem 2 (Multi-prototype representation gain)**.** *For $M_c > 1$,*

$$E_1^\star(c) = Z_c D_c^\star, \qquad Z_c = 1 - \sum_k w_{k,c}^2,$$

and the optimal unit-norm single-prototype error lies in $[E_1^\star(c), 2E_1^\star(c)]$. For every $M \geq 1$,

$$0 \leq E_1^\star(c) - E_M^\star(c) \leq Z_c D_c^\star.$$

The gain is strictly positive under the separated-group condition of Corollary 20 when sufficient representatives are retained, and monotonicity persists for unit-norm representatives.

Proposition 3 (Score-to-operation monotonicity)**.** *The realized bank capacity*

$$M(s) = \min\left\{M_c,\, 1 + \left\lfloor s(M_{\max} - 1) + \tfrac{1}{2}\right\rfloor\right\}$$

is nondecreasing in s, while $E_{M(s)}^\star(c)$ is nonincreasing. With $\widetilde{w}_c = \mathrm{Norm}(\widetilde{W}_c)$, $\eta(s) = \eta_0(1-s)$, and $a_c(s) = \mathrm{Norm}((1-\eta(s))\widetilde{w}_c + \eta(s)\bar{p}_c)$,

$$\|a_c(s) - \widetilde{w}_c\|_2^2 \leq \frac{16\eta_0^2(1-s)^2}{(1-2\eta_0)^2}.$$

After restoring the classifier-row norm,

$$\|\bar{W}_c - \widetilde{W}_c\|_2^2 = \|\widetilde{W}_c\|_2^2 \|a_c(s) - \widetilde{w}_c\|_2^2,$$

so the corresponding displacement bound is also nonincreasing in s.

The implementation-level stationarity bound is given in Theorem 36.

5 Experiments

5.1 Setup

We evaluate label skew on CIFAR-100 and HAM10000 with Dirichlet concentration 0.5 and $K \in \{10, 20, 50\}$, domain shift on DomainNet with six domains, and context shift on NICO++ `track_1` with six contexts; the latter two use $K \in \{6, 30\}$. At $K = 6$, each domain or context forms one client; at $K = 30$, each is split uniformly across five clients. All methods use the same ImageNet-21K-pretrained ViT-B/16, rank-8 LoRA, client assignments, test partitions, and complete

Table 1: Main personalized federated learning (PFL) results (mean±std over three seeds, accuracy after five local fine-tuning epochs). Best in bold, strongest baseline underlined.

	CIFAR-100			HAM10000			DomainNet		NICO++	
Method	K=10	K=20	K=50	K=10	K=20	K=50	K=6	K=30	K=6	K=30
FedIT	83.34±0.94	86.00±0.78	83.61±1.17	71.19±1.61	77.55±1.35	75.19±1.84	75.41±0.57	75.34±0.68	91.22±0.36	90.84±0.42
FFA-LoRA	81.48±1.05	85.71±0.82	87.59±0.74	72.43±1.77	75.70±1.22	74.77±1.70	76.55±0.47	78.40±0.61	90.55±0.30	91.45±0.35
FedSA-LoRA	81.99±1.16	80.61±1.31	82.87±1.22	70.06±1.73	78.28±1.66	71.36±1.23	77.49±0.54	78.62±0.70	90.43±0.34	90.90±0.39
LoRA-FAIR	83.30±0.87	85.71±0.72	87.67±0.61	58.90±2.27	54.75±1.75	66.30±2.11	77.86±0.43	80.12±0.57	91.17±0.26	92.18±0.31
GGEUR	84.76±0.80	85.03±0.66	87.34±0.69	63.44±1.46	74.61±1.55	75.37±1.34	77.73±0.51	79.83±0.51	91.11±0.24	92.14±0.36
HiLoRA	84.33±0.74	85.72±0.64	87.99±0.63	77.50±1.34	76.41±1.23	77.25±1.39	77.30±0.46	80.20±0.54	91.11±0.31	92.45±0.30
FEDNASC	**86.23±0.62**	**87.91±0.65**	**90.01±0.58**	**79.45±0.83**	**80.72±1.01**	**78.64±0.47**	**79.85±0.43**	**82.18±0.53**	**93.14±0.34**	**93.91±0.27**

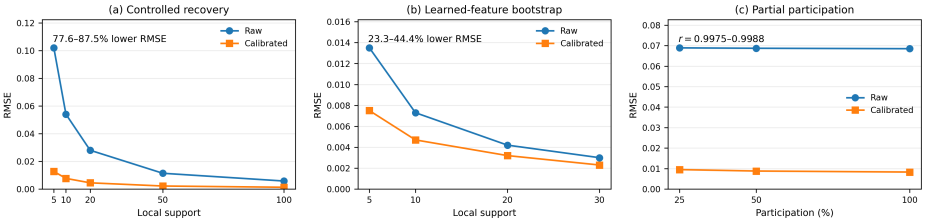


Figure 3: Noise-calibration diagnostics. Panels (a)–(b) compare raw and calibrated RMSE under controlled recovery and learned-feature bootstrap, while panel (c) evaluates client subsampling.

local training partitions. Label-skew experiments use 50 rounds with three local epochs, while domain and context shift use 20 rounds with two. The primary PFL evaluation follows five local personalization epochs from the final federated checkpoint. This stage is applied consistently across methods to evaluate client-specific adaptation rather than to extend federated optimization. We compare FedIT (Zhang et al., 2024c), FFA-LoRA (Sun et al., 2024), FedSA-LoRA (Guo et al., 2025), LoRA-FAIR (Bian et al., 2025), GGEUR (Ma et al., 2025), and HiLoRA (Peng et al., 2026), using source-recommended parameters when available and the same pilot-selection protocol otherwise, with no test-set tuning. Appendix B gives the complete protocol and additional results.

5.2 Main Results

Table 1 shows that FEDNASC achieves the highest mean accuracy in all ten settings, improving over the strongest baseline by 1.39–2.44 percentage points. Under label skew, the gains on CIFAR-100 are 1.47, 1.91, and 2.02 points for $K = 10, 20, 50$, respectively, while HAM10000 shows improvements of 1.95, 2.44, and 1.39 points. The largest gain occurs on HAM10000 at $K = 20$, while CIFAR-100 shows positive margins across all three scaling settings. Under structured distribution shifts, FEDNASC also maintains consistent gains of 1.99 and 1.98 points on DomainNet and 1.92 and 1.46 points on NICO++. Notably, the strongest competing method varies across datasets and federation sizes, whereas FEDNASC remains best throughout, indicating that its advantage is not tied to a single heterogeneity regime or baselines.

5.3 Ablation and Diagnostic Analysis

Noise calibration. We first directly evaluate the calibrated quantity in Theorem 1. As shown in Figure 3, matched subtraction reduces RMSE against true dispersion by 77.6–87.5% under controlled recovery (panel a) and by 23.3–44.4% under learned-feature bootstrap (panel b). Under client subsampling (panel c), calibrated RMSE remains 0.0083–0.0095, with correlation 0.9975–0.9988 against the exact active-set target. Together, these results show that the calibrated excess tracks the target dispersion under controlled, learned-feature, and varying-participation settings.

Functional components. Table 2 shows distinct roles for the controller components. Removing noise calibration causes the largest drop on CIFAR-100 (1.09 points), whereas collapsing the bank to a single prototype has the largest effect on DomainNet (1.22 points). This contrast is consistent

Table 2: Functional ablations on CIFAR-100 (K=10) and DomainNet (K=6). Δ is the change relative to the complete method. The largest drop per dataset is bold.

	CIFAR-100		DomainNet	
Functional variant	Accuracy	Δ	Accuracy	Δ
Full FEDNASC	86.23±0.62	—	79.85±0.43	—
Noise calibration				
w/o noise floor ($\nu_c \equiv 0$)	85.14±0.85	**−1.09**	79.22±0.62	−0.63
Adaptive bank construction				
Single prototype ($M_{\max}$=1)	85.56±0.76	−0.67	78.63±0.66	**−1.22**
Fixed full bank (no score control)	85.47±0.82	−0.76	79.02±0.61	−0.83
Semantic training				
w/o semantic loss (λ=0)	85.57±0.72	−0.66	78.96±0.57	−0.89
w/o anchoring (η_0=0)	85.82±0.79	−0.41	79.66±0.46	−0.19
Readout				
Linear-only ($\alpha_k^\star$=1)	85.97±0.73	−0.26	79.02±0.50	−0.83

Table 3: Participation-rate ablation on CIFAR-100 (K=50) at $\rho \in \{0.2, 0.5\}$, before and after five-epoch fine-tuning (mean±std, three seeds). Best in bold, strongest baseline underlined.

	$\rho = 0.2$		$\rho = 0.5$	
Method	Before FT	After FT	Before FT	After FT
FedIT	87.00±0.53	<u>83.00±0.65</u>	87.52±0.22	84.47±0.27
FFA-LoRA	86.02±0.61	81.05±0.09	86.83±0.91	84.50±0.28
FedSA-LoRA	83.25±0.45	78.55±0.12	84.81±1.21	81.98±0.95
LoRA-FAIR	<u>87.25±0.78</u>	82.90±0.68	87.44±0.30	85.29±0.13
GGEUR	86.52±0.63	82.67±0.96	87.67±0.62	85.47±0.27
HiLoRA	86.92±0.74	82.88±1.14	<u>87.78±0.03</u>	<u>85.90±0.10</u>
FEDNASC	**88.45±0.21**	**84.72±0.27**	**89.20±0.21**	**86.58±0.10**

with calibration being especially important under sparse label support, while multiple representatives are more valuable when persistent domain-specific class structure must be preserved. The fixed full bank also underperforms the adaptive bank on both datasets, showing that the gain comes from score-controlled capacity rather than simply retaining more prototypes. Semantic guidance and prototype-readout fusion provide further gains; the latter uses a second encoder forward when active, while the compute-matched linear-only variant still outperforms the strongest baseline on both datasets. Classifier anchoring produces smaller but consistent gains and acts as a secondary stabilization mechanism.

Client participation. We further vary the client participation rate on CIFAR-100 with K=50. The calibration diagnostic above shows that the estimator continues to track the exact active-set target under client subsampling. Table 3 evaluates the corresponding end-to-end performance. FEDNASC achieves the highest mean in all four reduced- participation conditions. Its margins over the strongest baseline are 1.20 and 1.72 points before and after fine-tuning at $\rho = 0.2$, and 1.42 and 0.68 points at $\rho = 0.5$. The corresponding standard deviations remain 0.10–0.27, showing that the advantage is maintained across seeds as participation decreases. Additional diagnostics and hyperparameter sweeps are provided in Appendix B.

6 CONCLUSION

FEDNASC calibrates cross-client prototype dispersion against finite-sample instability and uses the resulting score to control bank capacity, anchoring, semantic guidance, and readout fusion. Across four benchmarks and ten settings, it achieves the highest mean accuracy, exceeding the strongest baseline by 1.39–2.44 points. Ablations show that noise calibration is most important under sparse label skew, whereas adaptive capacity matters most under persistent domain shift. Together with theory and estimator diagnostics, these results support noise-aware semantic control. Two limitations merit further study. The current score does not disentangle data heterogeneity from encoder drift, and prototype-assisted inference requires an additional encoder pass. Future work will investigate drift-aware calibration and more efficient semantic readout.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used to assist with language editing, LaTeX formatting, and selected coding tasks, including code editing, debugging, and implementation refinement. They also provided limited assistance in checking intermediate mathematical derivations and reviewing proofs. All AI-assisted mathematical and coding work was reviewed and verified by the authors, who take full responsibility for the content of this work.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Fedalt: Federated fine-tuning through adaptive local training with rest-of-world lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19728–19736, 2026.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7314–7322, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Boyu Fan, Xiang Su, Sasu Tarkoma, and Pan Hui. Helora: Lora-heterogeneous federated fine-tuning for foundation models. *ACM Transactions on Internet Technology*, 25(2):1–22, 2025.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, volume 2025, pp. 99003–99027, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Kaili Jin, Li Xu, Xiaoding Wang, Sun-Yuan Hsieh, Jie Wu, and Limei Lin. Fedcpd: Personalized federated learning with prototype-enhanced representation and memory distillation. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

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

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

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.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021c.

Ying Li, Xingwei Wang, Rongfei Zeng, Praveen Kumar Donta, Ilir Murturi, Min Huang, and Schahram Dustdar. Federated domain generalization: A survey. *Proceedings of the IEEE*, 2025.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

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, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Shiwei Lu, Yuhang He, Jiashuo Li, Qiang Wang, and Yihong Gong. Gfpl: Generative federated prototype learning for resource-constrained and data-imbalanced vision task. *arXiv preprint arXiv:2602.21873*, 2026.

Yanbiao Ma, Wei Dai, Wenke Huang, and Jiayi Chen. Geometric knowledge-guided localized global distribution alignment for federated learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20958–20968, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

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.

Zihao Peng, Nan Zou, Jiandian Zeng, Guo Li, Ke Chen, Boyuan Li, and Tian Wang. Hilora: Hierarchical low-rank adaptation for personalized federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31746–31757, 2026.

Hariharan Ramesh and Jyotikrishna Dass. Spectral transformation for layer-wise global rank discovery in federated lora for vision transformers. *arXiv preprint arXiv:2607.21074*, 2026.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of foundation models. *arXiv preprint arXiv:2410.09432*, 2024.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*, 2024.

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, 2022a.

Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022b.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 180161, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Lei Wang, Jieming Bian, Letian Zhang, Chen Chen, and Jie Xu. Taming cross-domain representation variance in federated prototype learning with heterogeneous data domains. *Advances in Neural Information Processing Systems*, 37:88348–88372, 2024a.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024b.

Jianqiao Zhang, Caifeng Shan, and Jungong Han. Fedgmkd: An efficient prototype federated learning framework through knowledge distillation and discrepancy-aware aggregation. *Advances in Neural Information Processing Systems*, 37:118326–118356, 2024a.

Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16768–16776, 2024b.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE, 2024c.

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. NICO++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16036–16047, 2023.

Yong Zhang, Feng Liang, Guanghu Yuan, Min Yang, Chengming Li, and Xiping Hu. Fedpall: prototype-based adversarial and collaborative learning for federated learning with feature drift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3111–3120, 2025.

A Training and Inference Algorithms

A.1 Deferred Definitions

The main text gives the core operations of FedNASC. We collect the remaining implementation definitions here. Normalization is

$$\mathrm{Norm}(v) = \begin{cases} v/\|v\|_2, & \|v\|_2 > \epsilon, \\ 0, & \text{otherwise}, \end{cases} \qquad \epsilon = 10^{-12}, \tag{19}$$

and a zero output is marked invalid. The corresponding semantic operation is then skipped for that client and class in the current round.

The rebalancing step applied after Eq. (7) takes the reconstructed product $M = \bar{B}\bar{A}$ and forms its singular value decomposition $M = U\Sigma V^\top$. The leading r singular triplets are split symmetrically between the two factors,

$$\mathrm{Fact}_r(M) = (A, B), \qquad B = U_{:,1:r}\Sigma_{1:r}^{1/2}, \qquad A = \Sigma_{1:r}^{1/2} V_{:,1:r}^\top. \tag{20}$$

Since $M = \bar{B}\bar{A}$ has rank at most r, this refactorization preserves $BA = M$, balances the factor norms, and can be computed via thin QR factorizations followed by an $r \times r$ core SVD. The ridge reconstruction is constrained through $\bar{A}$, so its product is not generally the best rank-r approximation of the function-space target. We therefore record the per-layer relative reconstruction error

$$\frac{\|M - \Delta W_{\mathrm{avg}}\|_F}{\max\{\|\Delta W_{\mathrm{avg}}\|_F, \epsilon\}} \tag{21}$$

at each round, and the induced residual is carried through Lemma 35. Rebalancing is applied from round $t_{\mathrm{warm}} = 1$ onward.

In Eq. (7), the effective ridge weight is

$$\lambda_{\mathrm{r}} = 10^{-4} \frac{\mathrm{tr}(\bar{A}\bar{A}^\top)}{r}, \tag{22}$$

which scales the regularization with the overall magnitude of the LoRA factors.

The broadcast training gate is

$$g_c^{t+1} = \gamma + (1-\gamma) s_c^{t+1}, \qquad \gamma \in [0, 1), \tag{23}$$

and the upload package of client k in round t is

$$\mathcal{U}_k^t = \left\{ \Delta_k^{t+1}, \phi_k^{t+1} \right\} \cup \left\{ (p_{k,c}^t, n_{k,c}, \widehat{\nu}_{k,c}^t) : n_{k,c} \geq n_{\min} \right\}. \tag{24}$$

The frozen final reference encoder used by the semantic refresh and prototype readout is

$$\widetilde{f}_{\mathrm{ref}}^\star(x) = \mathrm{Norm}\left(F_{\theta_0, \bar{\Delta}^T, \bar{\psi}^T}(x) \right). \tag{25}$$

Writing $\mathcal{C}_{\mathcal{B}}^\star$ for the classes covered by the final bank and $u_{k,c}^\star$ for the class-level log-mean-exp logit of Eq. (2) evaluated under Eq. (25), the coverage-safe prototype readout rescales the prototype posterior by the linear probability mass assigned to covered classes and preserves the linear probability of each uncovered class,

$$q_{\mathrm{proto},k,c}(x) = \begin{cases} \left(\sum_{c' \in \mathcal{C}_{\mathcal{B}}^\star} q_{\mathrm{lin},k,c'}(x) \right) \dfrac{\exp(u_{k,c}^\star(x))}{\sum_{c' \in \mathcal{C}_{\mathcal{B}}^\star} \exp(u_{k,c'}^\star(x))}, & c \in \mathcal{C}_{\mathcal{B}}^\star, \\ q_{\mathrm{lin},k,c}(x), & \text{otherwise}. \end{cases} \tag{26}$$

The resulting distribution sums to one and reduces to the linear readout when the final bank is empty.

Algorithm 1 FEDNASC federated training

Require: Initial state $(\bar{\Delta}^0, \bar{\phi}^0)$, rounds T, clients $[K]$
Initialize $d_c^0 = \nu_c^0 = s_c^0 = 0$, $g_c^0 = \gamma$, and $\mathcal{B}_c^0 = \emptyset$ for every class c
for $t = 0, \ldots, T-1$ **do**
  Sample $\mathcal{S}^t$ and broadcast $(\bar{\Delta}^t, \bar{\phi}^t, \mathcal{B}^t, \mathbf{g}^t)$
  for each $k \in \mathcal{S}^t$ in parallel **do**
    Initialize the local trainable state from $(\bar{\Delta}^t, \bar{\phi}^t)$
    Optimize Eq. (18) using Eqs. (1)–(3)
    Compute Eqs. (4) and (5)
    Upload $\mathcal{U}_k^t$ from Eq. (24)
  end for
  Aggregate each LoRA projection using Eqs. (6)–(7)
  if $t \geq t_{\text{warm}}$ **then**
    Refactorize each reconstructed LoRA update using Eq. (20)
  end if
  Average the classifier, bias, and adapted LayerNorm state
  for each class c **do**
    if $M_c^t \geq 2$ **then**
      Update $(d_c^{t+1}, \nu_c^{t+1}, s_c^{t+1})$ using Eqs. (8)–(12)
    else
      Retain $(d_c^{t+1}, \nu_c^{t+1}, s_c^{t+1}) = (d_c^t, \nu_c^t, s_c^t)$
    end if
    if $M_c^t \geq 1$ **then**
      Build $\mathcal{B}_c^{t+1}$ using Eqs. (13) and (14)
      Anchor the classifier using Eq. (15)
    else
      Retain $\mathcal{B}_c^{t+1} = \mathcal{B}_c^t$
    end if
    Set g_c^{t+1} using Eq. (23)
  end for
end for

Additional semantic-controller definitions. The scaling in Eq. (5) gives

$$\mathbb{E}[\widehat{\nu}_{k,c}] = 2\nu_{k,c} + O(\nu_{k,c}^{3/2})$$

under the fixed-map model. Consequently, $\frac{1}{2}(\widehat{\nu}_{i,c} + \widehat{\nu}_{j,c})$ matches the two endpoint prototype-noise contributions of a cross-client pair to leading order. Aggregating these pairwise contributions with the same weights as Eq. (9) gives the compact matched noise floor in Eq. (10).

Each bank representative also stores its unnormalized assigned support,

$$\pi_{c,m}^{t+1} = \sum_{k: z_{k,c}^{t+1} = m} n_{k,c}, \tag{27}$$

where $z_{k,c}^{t+1}$ is the representative assigned to $p_{k,c}^t$ by Eq. (14). These counts weight the within-class consensus in Eq. (15). No additional normalization is required because the resulting consensus direction is normalized.

A.2 PROCEDURES

Algorithms 1 and 2 separate the ordinary federated training rounds from the one-time final semantic refresh. The refresh uses the final global representation as a common reference and performs no gradient update.

B ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

This appendix gives the complete experimental protocol, detailed domain and context results, before-fine-tuning evaluation, and diagnostics omitted from the main text.

Algorithm 2 Final semantic refresh, local fine-tuning, and score-weighted inference

Require: Final global state $(\bar{\Delta}^T, \bar{\phi}^T)$ and all clients $[K]$
Freeze the final reference encoder in Eq. (25)
for each client $k \in [K]$ in parallel **do**
  Recompute class statistics on the local partition without gradients
  Store $(p^\star_{k,c}, n_{k,c}, \widehat{\nu}^\star_{k,c})$ for each valid class
  Upload only the refreshed class statistics
end for
Clear the training-round banks before rebuilding the final semantic state
for each class c **do**
  Form $\mathcal{R}^\star_c$ and set $M^\star_c = |\mathcal{R}^\star_c|$
  if $M^\star_c = 0$ **then**
    Set $d^\star_c = \nu^\star_c = s^\star_c = 0$ and $\mathcal{B}^\star_c = \emptyset$
  else if $M^\star_c = 1$ **then**
    Set $d^\star_c = \nu^\star_c = s^\star_c = 0$ and build one representative
  else
    Recompute $d^\star_c$ and $\nu^\star_c$ from $\mathcal{R}^\star_c$ using the instantaneous forms of Eqs. (8)–(10) without carrying over the training EMA
    Compute $s^\star_c$ from the refreshed statistics using the score form of Eq. (12)
    Compute the final bank capacity using the starred form of Eq. (13)
    Build $\mathcal{B}^\star_c$ using Eq. (14) without classifier anchoring
  end if
  Set $g^\star_c = \gamma + (1-\gamma)s^\star_c$
end for
for each client k **do**
  Reinitialize the local optimizer and learning-rate state and set all momentum buffers to zero
  Starting from $(\bar{\Delta}^T, \bar{\phi}^T)$, perform five local fine-tuning epochs using Eq. (18) with $(\mathcal{B}^\star, \mathbf{g}^\star)$ fixed
  Form the adapted linear readout and the fixed final-reference prototype readout
  Form $\widehat{\alpha}_k$ from the refreshed scores using Eq. (16)
  if $\mathcal{B}^\star = \emptyset$ **then**
    Set $\alpha^\star_k = 1$
  else
    Select $\alpha^\star_k$ using Eq. (17)
  end if
  Predict with $\arg\max_c q^{(\alpha^\star_k)}_{k,c}(x)$
end for

B.1 Complete Experimental Protocol

Datasets and heterogeneity settings. We study four forms of heterogeneity. *CIFAR-100* (Krizhevsky et al., 2009) uses a $\mathrm{Dir}(0.5)$ partition over class proportions to create label skew. *HAM10000* (Tschandl et al., 2018) combines the same client partition with its natural seven-class long-tail distribution; images are first grouped by `lesion_id` and split 80:20 at lesion level, so no lesion appears in both train and test sets. *DomainNet* (Peng et al., 2019) uses a fixed 100-class subset and the six official domains (Clipart, Infograph, Painting, Quickdraw, Real, Sketch). *NICO++* (Zhang et al., 2023) uses the six `track_1` contexts (autumn, dim, grass, outdoor, rock, water). For DomainNet and NICO++, $K = 6$ assigns one domain/context to each client, whereas $K = 30$ splits each domain/context uniformly across five clients without mixing shift categories. All class lists, file identities, client assignments, test indices, and seeds are fixed before training and reused across methods.

Model and federated protocol. All methods use an ImageNet-21K-pretrained ViT-B/16 backbone (Dosovitskiy et al., 2020), identical preprocessing, client assignments, and test partitions. FedNASC trains rank-8 LoRA modules (scaling 1.0) on the fused query–key–value and output projections of each attention block, selected LayerNorm affine parameters, and the linear classifier. LoRA baselines use the same backbone and rank while retaining the trainable, communicated, and

inference components of their original designs; GGEUR retains its geometry-guided objective under the common backbone. All methods therefore share the backbone, data protocol, and training budget while preserving their method-specific adaptation designs.

The main benchmarks use full client participation. CIFAR-100 and HAM10000 use $T = 50$ rounds with $E_{\text{ep}} = 3$ local epochs, while DomainNet and NICO++ use $T = 20$ rounds with $E_{\text{ep}} = 2$. For label skew, $K \in \{10, 20, 50\}$; training fractions of 10%, 20%, and 50% are used, respectively, keeping approximate per-client volume comparable while increasing the total training pool with K. These comparisons therefore study federation scaling at approximately fixed per-client data, not client count at fixed total sample size. Local optimization uses SGD with momentum 0.9, weight decay 10^{-4}, gradient clipping at 5.0, batch size 32, learning rate 0.01 for LoRA/classifier parameters, and 0.001 for LayerNorm parameters. The same optimizer, schedule, round budget, and local epoch budget are used for every method.

Implementation constants and schedules. We use $n_{\min} = 5$, $\epsilon = 10^{-12}$, and 20 weighted spherical k-means iterations per bank update. With zero-indexed rounds $t = 0, \dots, T-1$, the scarcity ramp is

$$\rho_t = \min\left\{1, \max\left\{0, \frac{t-1}{3}\right\}\right\}, \tag{28}$$

so $\rho_0 = \rho_1 = 0$, $\rho_2 = 1/3$, $\rho_3 = 2/3$, and $\rho_t = 1$ for $t \geq 4$. Weighted spherical k-means uses support-weighted k-means++ initialization, normalized support-weighted cluster centers, and reinitializes an empty cluster to the prototype with largest current weighted angular loss; ties are resolved by client index.

The shared controller constants are

$$\begin{aligned} \gamma = 0.1, \quad n_0 = 50, \qquad \kappa = 0.25, \\ \eta_0 = 0.1, \quad \beta = 0.65, \quad \alpha_{\min} = 0.35. \end{aligned}$$

These values are fixed across datasets. In particular, $n_{\min} = 5$ ensures that both split halves contain at least two examples for every uploaded class, $\gamma > 0$ prevents a covered semantic term from being switched off completely, and $\eta_0 < 1/2$ ensures $1 - 2\eta_0 > 0$ in the anchoring normalization bound.

Evaluation and baselines. Every reported run uses the final scheduled communication round. Each participating client then starts from the final federated checkpoint and performs five local fine-tuning epochs. The optimizer and momentum state are reinitialized, while the refreshed semantic state remains fixed and no further server update is performed during this stage. $\alpha_k^\star$ is reselected on the client's training partition before evaluation on its unchanged test split. We also report accuracy immediately before local fine-tuning using the same final checkpoint and refreshed bank. All results are mean$\pm$sample standard deviation over seeds $0, 1, 2$.

We compare FedIT (Zhang et al., 2024c), FFA-LoRA (Sun et al., 2024), FedSA-LoRA (Guo et al., 2025), LoRA-FAIR (Bian et al., 2025), GGEUR (Ma et al., 2025), and HiLoRA (Peng et al., 2026) in one codebase with the same backbone, partitions, optimizer, preprocessing, seeds, and training budgets. Baseline-specific parameters use source-recommended settings when available; for uncovered dataset or configuration settings, parameters are selected under the same pilot protocol as FEDNASC, without test-set tuning. GGEUR retains its geometry-guided embedding and prototype-transfer objectives under the common backbone. Baselines use their prescribed inference rules, whereas FEDNASC uses the proposed prototype readout and uploaded semantic statistics.

Split-half rule, sensitivity, and ablations. The statistic in Eq. (5) is computed only for $n_{k,c} \geq 5$; all examples are retained and odd supports use the generalized coefficient $\chi_{k,c}$. The shared FEDNASC configuration $(\lambda, \tau, M_{\max}, \mu_{\text{ema}}) = (0.5, 0.1, 8, 0.7)$ was fixed under this pilot protocol and reused across datasets. The one-factor sweeps were separate sensitivity runs assessing relative changes around the shared default.

The functional ablations change one operation at a time. *w/o noise floor* sets $\nu_c^t = 0$; *single prototype* sets $M_{\max} = 1$; *fixed full bank* uses $\min\{M_c^t, M_{\max}\}$; *w/o semantic loss* sets $\lambda = 0$; *w/o anchoring* sets $\eta_0 = 0$; and *linear-only* fixes $\alpha_k^\star = 1$.

Method	Avg	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
FedIT	75.41±0.57	79.64±1.13	54.69±1.18	78.39±1.03	73.90±1.01	88.20±1.36	77.66±1.31
FFA-LoRA	76.55±0.47	81.55±1.06	55.16±0.98	79.55±1.04	74.73±1.01	88.99±1.12	79.31±1.17
FedSA-LoRA	77.49±0.54	81.74±1.23	55.14±0.92	80.16±0.95	77.99±1.26	89.44±1.10	80.48±1.02
LoRA-FAIR	77.86±0.43	83.35±0.75	55.85±0.85	80.53±1.06	77.15±0.97	89.18±0.94	81.09±1.20
GGEUR	77.73±0.51	82.84±1.01	56.03±0.94	80.82±0.83	76.76±0.85	89.32±0.89	80.59±0.90
HiLoRA	77.30±0.46	82.15±0.97	55.84±0.88	79.64±0.96	76.92±0.66	89.33±1.10	79.91±0.82
FEDNASC (ours)	**79.85±0.43**	**85.44±0.68**	**58.99±0.78**	**82.23±0.74**	**79.18±0.65**	**90.86±0.81**	**82.40±0.73**

Table 4: Client-average PFL accuracy (%) after five epochs of personalized fine-tuning under domain shift on DomainNet (100 classes) with $K = 6$, where each domain is assigned to one client. Results are mean±std over three seeds. The Avg value is formed within each seed across the six domains, and its std is then computed across the three seed-level averages. The best and second-best mean results are marked in bold and with an underline, respectively.

Method	Avg	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
FedIT	75.34±0.68	78.64±1.31	56.76±1.20	78.24±1.10	74.34±1.22	89.96±0.95	74.12±1.22
FFA-LoRA	78.40±0.61	82.83±0.82	58.74±1.04	80.79±0.82	78.48±0.97	90.10±0.99	79.43±1.06
FedSA-LoRA	78.62±0.70	83.73±1.13	58.46±1.02	81.54±0.90	78.20±0.76	90.32±0.83	79.49±1.21
LoRA-FAIR	80.12±0.57	85.09±0.76	62.26±0.97	83.21±1.03	78.76±1.07	91.05±0.75	80.32±1.10
GGEUR	79.83±0.51	84.22±0.70	61.08±0.70	82.81±0.70	78.79±0.89	90.75±0.92	81.33±0.80
HiLoRA	80.20±0.54	84.84±0.81	61.88±0.76	82.79±0.90	79.61±0.86	90.93±0.91	81.14±0.79
FEDNASC (ours)	**82.18±0.53**	**86.93±0.64**	**63.91±0.63**	**85.26±0.71**	**81.63±0.68**	**92.40±0.76**	**82.95±0.72**

Table 5: Client-average PFL accuracy (%) after five epochs of personalized fine-tuning under domain shift on DomainNet (100 classes) with $K = 30$, where each domain is split uniformly across five clients. Results are mean±std over three seeds. The Avg value is formed within each seed across the six domains, and its std is then computed across the three seed-level averages. The best and second-best mean results are marked in bold and with an underline, respectively.

B.2 Per-Domain, Per-Context, and Before Local Fine-Tuning Results

Tables 4–7 show that FEDNASC has the highest mean in every reported DomainNet domain and NICO++ context at both federation scales. On DomainNet with $K = 6$, its largest margins over the strongest baseline are 2.96 points on Infograph and 2.09 on Clipart, while the margin on Real is 1.42. At $K = 30$, the overall margins are 1.98 points on DomainNet and 1.46 on NICO++.

Accuracy before and after personalized fine-tuning. The main tables report accuracy after five local fine-tuning epochs. The next two tables report the same final federated states before this stage.

FEDNASC has the highest before-fine-tuning mean in all ten configurations. Its margins over the runner-up are 1.81, 2.07, and 2.00 points on CIFAR-100; 1.52, 1.83, and 1.53 on HAM10000; 1.61 and 1.85 on DomainNet; and 1.44 and 2.17 on NICO++. Fine-tuning reduces FEDNASC in nine of ten settings by 0.02–2.44 points; the only increase is +0.26 on HAM10000 at $K = 20$. These results distinguish the performance of the final federated checkpoint from that obtained after client-specific adaptation.

B.3 Extended Estimator and Controller Diagnostics

Controlled recovery of population dispersion. To validate the quantity in Theorem 14 directly, we use a hyperspherical experiment with 20 clients and $d = 128$. Half of the clients belong to each of two equally weighted population modes separated by $\{0°, 20°, 40°, 60°\}$. For a unit direction μ,

$$\Pi_\mu = I - \mu\mu^\top, \qquad \xi \sim \mathcal{N}\left(0, \frac{0.8^2}{d-1}\Pi_\mu\right), \qquad z = \mathrm{Norm}(\mu + \xi).$$

Here 0.8 is the root-mean-square tangent perturbation norm. The largest perturbations extend beyond the local regime of Assumption 8, making this experiment an empirical robustness diagnostic in addition to the local-regime analysis. We vary $n_{k,c} \in \{5, 10, 20, 50, 100\}$ and repeat each condition 300 times.

Method	Avg	Autumn	Dim	Grass	Outdoor	Rock	Water
FedIT	91.22±0.36	92.43±0.55	90.19±0.63	93.80±0.70	90.09±0.74	90.27±0.42	90.52±0.53
FFA-LoRA	90.55±0.30	92.54±0.47	88.33±0.56	93.77±0.53	89.14±0.54	89.55±0.58	89.95±0.43
FedSA-LoRA	90.43±0.34	91.44±0.44	89.75±0.36	93.02±0.60	88.74±0.49	89.60±0.56	90.00±0.58
LoRA-FAIR	91.17±0.26	92.48±0.48	90.27±0.45	93.85±0.42	89.79±0.41	90.13±0.57	90.49±0.45
GGEUR	91.11±0.24	92.13±0.49	90.07±0.27	93.70±0.37	89.64±0.47	90.62±0.26	90.52±0.41
HiLoRA	91.11±0.31	92.16±0.32	90.19±0.39	93.73±0.44	90.18±0.39	89.96±0.30	90.44±0.43
FEDNASC (ours)	**93.14±0.34**	**94.03±0.43**	**92.38±0.24**	**95.38±0.48**	**92.62±0.34**	**91.89±0.37**	**92.54±0.36**

Table 6: Client-average PFL accuracy (%) after five epochs of personalized fine-tuning under context shift on NICO++ (`track_1`) with $K = 6$, where each context is assigned to one client. Results are mean±std over three seeds. The Avg value is formed within each seed across the six contexts, and its std is then computed across the three seed-level averages. The best and second-best mean results are marked in bold and with an underline, respectively.

Method	Avg	Autumn	Dim	Grass	Outdoor	Rock	Water
FedIT	90.84±0.42	92.57±0.47	89.34±0.69	93.63±0.60	89.56±0.59	89.82±0.50	90.10±0.66
FFA-LoRA	91.45±0.35	92.74±0.67	90.08±0.30	93.90±0.61	90.84±0.65	90.37±0.57	90.77±0.33
FedSA-LoRA	90.90±0.39	92.44±0.56	89.46±0.44	93.28±0.50	90.16±0.34	90.09±0.60	89.95±0.46
LoRA-FAIR	92.18±0.31	93.78±0.31	90.62±0.58	94.36±0.37	91.50±0.36	91.31±0.25	91.48±0.45
GGEUR	92.14±0.36	93.73±0.30	90.55±0.57	94.38±0.47	91.44±0.65	91.26±0.50	91.47±0.40
HiLoRA	92.45±0.30	93.34±0.51	91.04±0.35	94.40±0.46	92.24±0.50	91.86±0.45	91.82±0.40
FEDNASC (ours)	**93.91±0.27**	**95.02±0.36**	**92.79±0.27**	**95.93±0.33**	**93.89±0.35**	**92.93±0.29**	**92.90±0.28**

Table 7: Client-average PFL accuracy (%) after five epochs of personalized fine-tuning under context shift on NICO++ (`track_1`) with $K = 30$, where each context is split uniformly across five clients. Results are mean±std over three seeds. The Avg value is formed within each seed across the six contexts, and its std is then computed across the three seed-level averages. The best and second-best mean results are marked in bold and with an underline, respectively.

At zero semantic separation, raw dispersion is 0.1130 at support 5 and 0.0063 at support 100, despite the true dispersion being zero; matched subtraction reduces the corresponding calibrated excess to 0.0127 and 0.0001. Across all separations, calibration reduces RMSE by 77.6–87.5%.

Learned-feature bootstrap on real images. We next test the estimator on non-Gaussian learned features using the public handwritten-digit images distributed with scikit-learn. A small convolutional classifier is trained on five deterministic appearance domains: the original images, two $\pm 18^\circ$ rotations, Gaussian blur, and a subpixel translation. The architecture, training setup, transformations, split, and seed are fixed, and normalized 64-dimensional penultimate features are extracted. For each class and domain, the complete held-out feature pool defines a high-support finite-sample empirical reference. We bootstrap supports $n \in \{5, 10, 20, 30\}$ with replacement from the same pool and repeat each support/class condition 200 times. Errors are measured relative to this empirical reference.

Calibration reduces RMSE by 44.4%, 35.6%, 23.8%, and 23.3% at supports 5, 10, 20, and 30, respectively, while calibrated bias remains within 2.3×10^{-4}. The largest reduction at low support is consistent with the intended finite-sample correction.

Estimator robustness under client subsampling. A controlled subsampling diagnostic uses 40 clients, $\rho \in \{0.25, 0.5, 1.0\}$, supports $n_{k,c} \in \{5, 10, 20\}$, the same four semantic separations, and 300 repetitions per condition. The exact population dispersion is recomputed for every sampled active set.

Calibrated RMSE changes from 0.0083 under full participation to 0.0095 at 25% participation, while correlation with the exact active-set target remains above 0.997. The estimator therefore remains accurate across the tested participation rates. End-to-end partial-participation results are reported in the main text.

Sensitivity of the exposed configuration parameters. We vary one parameter at a time on CIFAR-100 ($K = 10$) and DomainNet ($K = 6$): $\lambda \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$, $\tau \in$

Method	CIFAR-100			HAM10000		
	$K = 10$	$K = 20$	$K = 50$	$K = 10$	$K = 20$	$K = 50$
FedIT	86.64 ± 0.80	87.65 ± 0.94	84.29 ± 1.04	75.14 ± 1.82	78.42 ± 1.53	76.01 ± 1.58
FFA-LoRA	85.27 ± 1.14	87.22 ± 0.75	88.62 ± 0.91	75.07 ± 1.43	77.41 ± 1.14	75.84 ± 1.92
FedSA-LoRA	84.54 ± 0.85	83.28 ± 1.12	83.24 ± 1.31	75.55 ± 1.89	78.62 ± 1.45	73.85 ± 1.07
LoRA-FAIR	86.86 ± 0.98	87.10 ± 0.67	88.50 ± 0.75	68.71 ± 2.00	68.04 ± 1.99	62.60 ± 1.82
GGEUR	86.28 ± 0.63	86.90 ± 0.54	88.33 ± 0.84	75.29 ± 1.23	77.42 ± 1.74	76.85 ± 1.54
HiLoRA	86.82 ± 0.53	87.63 ± 0.39	88.52 ± 0.57	78.56 ± 1.19	78.63 ± 1.18	78.21 ± 0.98
FEDNASC (ours)	$\mathbf{88.67 \pm 0.44}$	$\mathbf{89.72 \pm 0.72}$	$\mathbf{90.62 \pm 0.42}$	$\mathbf{80.08 \pm 0.94}$	$\mathbf{80.46 \pm 0.99}$	$\mathbf{79.74 \pm 1.29}$

Table 8: Client-average accuracy (%) immediately before the five-epoch personalized fine-tuning stage on CIFAR-100 and HAM10000. Results are reported as mean±sample standard deviation over three seeds.

Method	DomainNet		NICO++	
	$K = 6$	$K = 30$	$K = 6$	$K = 30$
FedIT	75.46 ± 0.69	74.97 ± 0.53	91.25 ± 0.29	90.86 ± 0.53
FFA-LoRA	77.36 ± 0.55	79.85 ± 0.79	90.77 ± 0.45	91.54 ± 0.26
FedSA-LoRA	78.53 ± 0.73	79.48 ± 0.62	91.05 ± 0.25	91.39 ± 0.47
LoRA-FAIR	78.43 ± 0.39	80.06 ± 0.73	91.16 ± 0.36	92.24 ± 0.25
GGEUR	78.31 ± 0.42	79.52 ± 0.40	91.00 ± 0.30	92.18 ± 0.43
HiLoRA	78.92 ± 0.53	80.35 ± 0.50	91.99 ± 0.35	92.27 ± 0.35
FEDNASC (ours)	$\mathbf{80.53 \pm 0.35}$	$\mathbf{82.20 \pm 0.55}$	$\mathbf{93.43 \pm 0.18}$	$\mathbf{94.44 \pm 0.22}$

Table 9: Client-average accuracy (%) immediately before the five-epoch personalized fine-tuning stage on DomainNet and NICO++. Results are reported as mean±sample standard deviation over three seeds.

$\{0.05, 0.075, 0.1, 0.15, 0.2\}$, $M_{\max} \in \{1, 2, 4, 8, 16\}$, and $\mu_{\text{ema}} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The shared default configuration is identical to that used in the corresponding main experiment, and the same default runs reported in Table 1 are used as the reference for all four sweeps. The default achieves the highest point estimate in all eight panels. Relative to this reference, the observed ranges are -1.76 to 0.00, -1.50 to 0.00, -1.41 to 0.00, and -1.39 to 0.00 on CIFAR-100, and -2.06 to 0.00, -1.84 to 0.00, -1.89 to 0.00, and -1.77 to 0.00 on DomainNet for λ, τ, $M_{\max}$, and μ_{ema}, respectively. Performance degrades more clearly toward the more extreme settings, while moderate deviations remain comparatively stable. For $M_{\max}$, deviations from the shared default reduce performance on both datasets.

Support	Raw RMSE	Calibrated RMSE	Reduction
5	0.1019	0.0127	87.5%
10	0.0540	0.0077	85.7%
20	0.0280	0.0045	83.9%
50	0.0114	0.0022	80.7%
100	0.0058	0.0013	77.6%

Table 10: Controlled recovery of population dispersion using the generalized split-half scale. RMSE is computed against the exact active-client $D_c^\star$ over four semantic separations and 300 repetitions.

Support	Raw bias	Cal. bias	Raw RMSE	Cal. RMSE
5	0.0113	0.0002	0.0135	0.0075
10	0.0058	0.0000	0.0073	0.0047
20	0.0029	0.0000	0.0042	0.0032
30	0.0019	0.0000	0.0030	0.0023

Table 11: Bootstrap validation on normalized features learned from real images. Errors are measured against high-support cross-domain feature dispersion.

ρ	Active	Raw RMSE	Cal. RMSE	Cal. corr.
0.25	10	0.0689	0.0095	0.9975
0.50	20	0.0687	0.0088	0.9983
1.00	40	0.0685	0.0083	0.9988

Table 12: Controlled estimator robustness under client subsampling. Correlation is between calibrated excess and exact active-set $D_c^\star$.

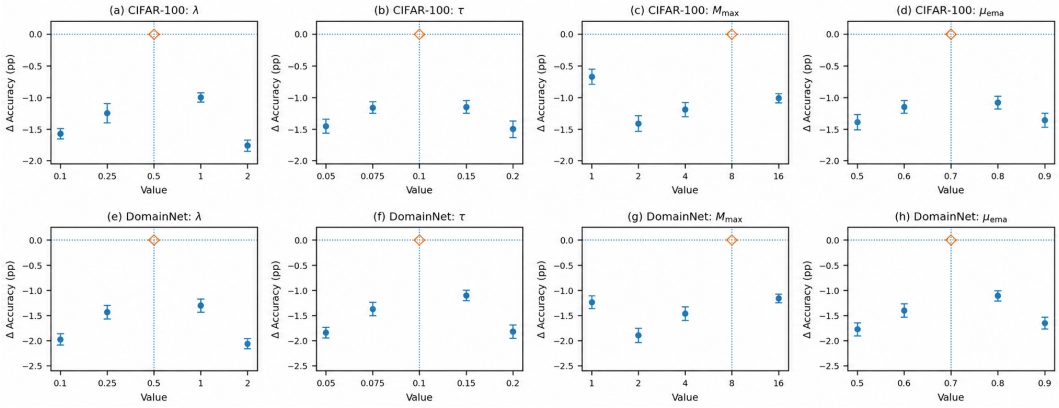


Figure 4: One-factor sensitivity on CIFAR-100 and DomainNet for λ, τ, $M_{\max}$, and μ_{ema}. Each point reports the change in after-fine-tuning accuracy relative to the shared default used in the corresponding Table 1 experiment over seeds 0, 1, and 2; error bars show one sample standard deviation of the paired differences. Hollow diamonds mark the shared defaults $(0.5, 0.1, 8, 0.7)$.

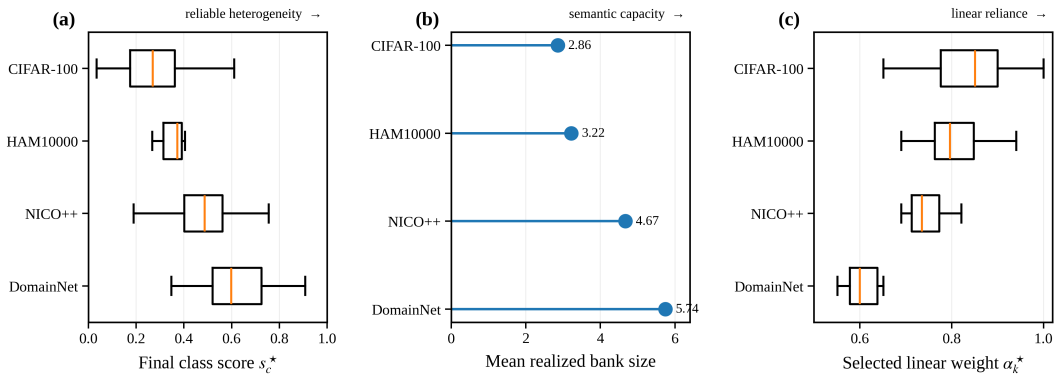


Figure 5: Final controller statistics under the main settings, pooled over seeds $0, 1, 2$: (a) refreshed class scores $s_c^\star$, (b) mean realized bank size, and (c) selected linear weights $\alpha_k^\star$. CIFAR-100 and HAM10000 use $K = 10$; DomainNet and NICO++ use $K = 6$.

Controller behavior across datasets. Figure 5 summarizes the final refreshed scores, realized bank sizes, and selected linear weights under the main settings. Mean realized bank sizes are 2.86, 3.22, 4.67, and 5.74 for CIFAR-100, HAM10000, NICO++, and DomainNet, respectively. Higher score distributions are accompanied by larger realized banks and lower selected linear weights.

Overhead. Each participating client uploads one d-dimensional normalized prototype, one support count, and one noise scalar per valid class, giving $O(|\mathcal{C}_k|d)$ client-to-server semantic communication. With $d = 768$, this is 770 scalars per valid class. CIFAR-100 clients at $K = 10$ average 42.1 valid classes, or 129.7KB in FP32 (5.8% of the LoRA/classifier/LayerNorm payload); DomainNet is approximately 308.0KB (13.8%).

The server also broadcasts the prototype bank and class-wise gate. Let $B_t = \sum_c |\mathcal{B}_c^t|$ denote the number of stored representatives in round t. Ignoring serialization overhead, the semantic downlink contains $B_t(d + 1) + C$ scalars per selected client—one d-vector and one stored support per representative, plus one gate value per class—and is therefore $O(B_t d + C)$, with the worst-case bound $O(CM_{\max}d)$. The same form applies to the one-time refreshed final bank. Because B_t varies across rounds and classes, the downlink depends on the realized bank size.

Semantic statistics require one no-gradient local pass after optimization, and the final refresh adds one further pass. Server-side score estimation, bank construction, and anchoring take 0.21s per round on CIFAR-100 and 0.47s on DomainNet (1.2% and 1.6% of round wall-clock time, excluding client feature extraction).

At inference, $\alpha_k^\star = 1$ uses one encoder forward pass. When $\alpha_k^\star < 1$, the current implementation evaluates both the personalized and final-reference encoders, followed by $O(|\mathcal{B}^\star|d)$ prototype similarities. Intermediate activations are not reused, so backbone FLOPs are approximately doubled while the prototype path is active. The additional cost is computational rather than model storage, since both paths share the frozen backbone and differ only in lightweight adaptation states.

C Proofs

Relation to the main-text results. Theorem 1 corresponds to Theorem 14 below. Theorem 2 combines Theorem 17, Lemma 18, Theorem 19, and Corollary 20. Proposition 3 follows from Propositions 24 and 25.

The calibrated-dispersion and prototype-approximation analyses meet at $D_c^\star$. Theorem 14 shows that matched subtraction recovers this population dispersion up to the stated noise-dependent terms, while Theorem 17 gives the exact Euclidean identity $E_{\text{euc}}(c) = Z_c D_c^\star$. Theorem 19 establishes that optimal approximation error is nonincreasing with the number of retained prototypes, and Corollary 20 gives a separated-mode condition for strict Euclidean improvement. The score-to-operation

Parameter	Value	CIFAR-100	DomainNet
λ	0.1	84.66 ± 0.84	77.88 ± 0.62
	0.25	84.98 ± 0.67	78.42 ± 0.51
	$0.5^{\dagger}$	86.23 ± 0.62	79.85 ± 0.43
	1	85.23 ± 0.76	78.55 ± 0.41
	2	84.47 ± 0.91	77.79 ± 0.66
τ	0.05	84.78 ± 0.81	78.01 ± 0.61
	0.075	85.07 ± 0.74	78.48 ± 0.52
	$0.1^{\dagger}$	86.23 ± 0.62	79.85 ± 0.43
	0.15	85.08 ± 0.65	78.75 ± 0.49
	0.2	84.73 ± 0.83	78.03 ± 0.64
$M_{\max}$	1	85.56 ± 0.76	78.63 ± 0.66
	2	84.82 ± 0.77	77.96 ± 0.57
	4	85.04 ± 0.68	78.39 ± 0.49
	$8^{\dagger}$	86.23 ± 0.62	79.85 ± 0.43
	16	85.22 ± 0.75	78.69 ± 0.38
μ_{ema}	0.5	84.84 ± 0.85	78.08 ± 0.63
	0.6	85.08 ± 0.73	78.45 ± 0.52
	$0.7^{\dagger}$	86.23 ± 0.62	79.85 ± 0.43
	0.8	85.15 ± 0.63	78.74 ± 0.48
	0.9	84.87 ± 0.80	78.20 ± 0.58

Table 13: Absolute accuracies for the one-factor sensitivity runs corresponding to Figure 4. Each entry reports after-fine-tuning client-average accuracy as mean $\pm$ sample standard deviation over seeds $0, 1, 2$. Figure 4 reports paired changes relative to the shared default runs in Table 1.

results then establish the monotone capacity and anchoring mappings used in the main text, with separate sufficient conditions for semantic gating and inference fusion.

The convergence analysis complements these approximation results. Theorem 32 gives a fixed-H, plain-local-SGD reference bound. Lemma 33 and Theorem 36 transfer that reference to the implemented momentum/group-rate optimizer and ridge-based server map through explicit endpoint and functional-coordinate residuals.

The fixed-map and fitted-encoder cases are treated separately. When the encoder map is fixed independently of the current class sample, the randomized disjoint halves are independent. For the implemented fitted encoder, Remark 5 introduces the representation-reuse remainder $\delta_{\mathrm{rep},c}$, while Lemma 15 quantifies the additional lag introduced by exponential smoothing.

Throughout, constants hidden in $O(\cdot)$ are independent of the number of clients, communication rounds, and local class supports, although they may depend on the fixed moment and local-radius constants stated in the assumptions.

C.1 Preliminaries and notation

For a fixed round, condition on the server history before local adaptation. Let $(X, Y) \sim \mathcal{P}_k$ denote an independent draw from client k's underlying data distribution. After local optimization, freeze the realized feature map $\widetilde{f}_k$ during the no-gradient statistical pass and define the class-conditional feature law

$$\mathcal{P}_{k,c} = \mathcal{L}\Big(\widetilde{f}_k(X) \mid Y = c\Big) \tag{29}$$

on $\mathbb{S}^{d-1}$, where

$$\widetilde{f}_k(X) = \frac{f_k(X)}{\|f_k(X)\|_2} \in \mathbb{S}^{d-1}. \tag{30}$$

Throughout the theoretical analysis, every encoder feature and every population or empirical mean to which Norm is applied is assumed to have Euclidean norm strictly greater than ϵ. Consequently, the implementation in Eq. (19) coincides with exact normalization and every retained direction lies on $\mathbb{S}^{d-1}$.

The class-c local training examples are $\mathcal{D}_{k,c} = \{x : (x, y) \in \mathcal{D}_k,\, y = c\}$, with $n_{k,c} = |\mathcal{D}_{k,c}|$. This finite sample is distinct from the population law $\mathcal{P}_{k,c}$. The population class direction is

$$\mu_{k,c} = \mathrm{Norm}\big(\mathbb{E}_{z \sim \mathcal{P}_{k,c}}[z]\big)\,, \qquad \|\mu_{k,c}\|_2 = 1. \tag{31}$$

The empirical prototype uploaded by client k is

$$p_{k,c} = \mathrm{Norm}\left(\frac{1}{n_{k,c}} \sum_{x \in \mathcal{D}_{k,c}} \widetilde{f}_k(x)\right), \qquad \|p_{k,c}\|_2 = 1. \tag{32}$$

For unit vectors u, v, the identity

$$1 - \langle u, v \rangle = \frac{1}{2}\|u - v\|_2^2 \tag{33}$$

will be used repeatedly.

For a fixed round t and class c, define the valid client set

$$\mathcal{K}_c^t = \{k \in \mathcal{S}^t : n_{k,c} \geq n_{\min}\}, \qquad M_c^t = |\mathcal{K}_c^t|. \tag{34}$$

Whenever $M_c^t > 0$, define the reliability weights

$$w_{k,c}^t = \frac{\sqrt{n_{k,c}}}{\sum_{j \in \mathcal{K}_c^t} \sqrt{n_{j,c}}}, \qquad \sum_{k \in \mathcal{K}_c^t} w_{k,c}^t = 1. \tag{35}$$

For a fixed class and round, we suppress the superscript t and write M_c, $w_{k,c}$, $\mu_{k,c}$, $p_{k,c}$, and $\nu_{k,c}$. When both the class and round are fixed, we further write w_k, μ_k, p_k, ν_k, and $\mathcal{K}$ for readability. The reliability weights $w_{k,c}$ are used in the calibrated-dispersion analysis; the linear support weights $r_{k,c}$ used by the implemented bank are introduced in Section C.5.

Throughout the prototype analysis, the population directions and local supports are treated as fixed. All expectations in Sections C.2–C.5 are therefore taken over the prototype-sampling perturbations conditional on these quantities. In particular, the conditional cross-client independence in Eq. (50) implies the factorization in Eq. (51).

Lemma 4 (Independence generated by a random split)**.** *Let $Z_1, \ldots, Z_n$ be independent draws from a common distribution, and let Π be a uniformly random permutation independent of these draws. For fixed $m_a + m_b = n$, define*

$$\mathcal{Z}^{(a)} = (Z_{\Pi(1)}, \ldots, Z_{\Pi(m_a)}), \qquad \mathcal{Z}^{(b)} = (Z_{\Pi(m_a+1)}, \ldots, Z_{\Pi(n)}).$$

Then $\mathcal{Z}^{(a)}$ and $\mathcal{Z}^{(b)}$ are independent, with laws $\mathcal{P}^{m_a}$ and $\mathcal{P}^{m_b}$, respectively.

Proof. For every fixed permutation π, independence and identical distribution of $Z_1, \ldots, Z_n$ give

$$\mathcal{L}\big(Z_{\pi(1)}, \ldots, Z_{\pi(n)}\big) = \mathcal{P}^n = \mathcal{P}^{m_a} \otimes \mathcal{P}^{m_b}.$$

This joint law does not depend on π. Averaging it over the independent uniform permutation Π therefore leaves the same product law, which proves the claim.

Remark 5 (Direct correction for reuse after local fitting)**.** *Lemma 4 applies exactly to an encoder map that is fixed independently of the current class sample. The implemented statistical pass freezes its encoder, but that encoder was fitted on the same local training set. Conditioning on the fitted encoder alone does not restore independence. Instead, compare the implemented map f with a fixed-map reference f_0 that is independent of the current class split. Concretely, f_0 can be obtained by coupling the same initialization, optimizer schedule, non-c examples, and optimization randomness while replacing the class-c training examples by independent ghost draws; the current class split is then independent of f_0.*

Suppose that a deterministic replacement-stability argument for the fitted representation gives

$$\sup_x \|\widetilde{f}(x) - \widetilde{f}_0(x)\|_2 \leq \varepsilon_{\mathrm{rep}}, \qquad r_{\mathrm{mean}} > 0,$$

where r_{mean} lower-bounds the norm of every relevant empirical half/full normalized-feature mean and population normalized-feature mean under both maps. Averaging the pointwise bound shows that corresponding means differ by at most $\varepsilon_{\mathrm{rep}}$. The normalization inequality

$$\|\mathrm{Norm}(u) - \mathrm{Norm}(v)\|_2 \leq \frac{2\|u - v\|_2}{r_{\mathrm{mean}}}$$

therefore shows that every corresponding half prototype, full prototype, or population direction differs by at most

$$a_{\mathrm{rep}} = \frac{2\varepsilon_{\mathrm{rep}}}{r_{\mathrm{mean}}}.$$

Applied to client k and class c, this quantity is denoted $a_{\mathrm{rep},k,c}$. Since both prototypes are unit vectors,

$$\begin{aligned}&\left|\langle p^{(a)}, p^{(b)}\rangle - \langle p_0^{(a)}, p_0^{(b)}\rangle\right| \\ &\qquad \le \|p^{(a)} - p_0^{(a)}\|_2 + \|p^{(b)} - p_0^{(b)}\|_2 \le 2a_{\mathrm{rep}}.\end{aligned}$$

Because $\chi_{k,c} = 2m_{k,c}^{(a)} m_{k,c}^{(b)}/n_{k,c}^2 \le 1/2$, the implemented and reference split statistics satisfy the direct bound

$$|\widehat{\nu}_{k,c} - \widehat{\nu}_{k,c}^0| \le a_{\mathrm{rep},k,c}. \tag{36}$$

The same unit-vector inequality bounds the change of every raw or population pairwise angular distance by $2a_{\mathrm{rep},c}$, where $a_{\mathrm{rep},c} = \max_k a_{\mathrm{rep},k,c}$. The normalized pair weights sum to one, as do the compact noise-floor weights. Consequently,

$$\begin{aligned}&\left|\{(\widetilde{d}_c - \widetilde{\nu}_c) - D_c^\star\} - \{(\widetilde{d}_c^0 - \widetilde{\nu}_c^0) - D_c^{\star,0}\}\right| \\ &\qquad \le 2a_{\mathrm{rep},c} + a_{\mathrm{rep},c} + 2a_{\mathrm{rep},c} = 5a_{\mathrm{rep},c} \equiv \delta_{\mathrm{rep},c}.\end{aligned} \tag{37}$$

Thus the fixed-map independence argument requires no representation-reuse correction, whereas fitting and summarizing on the same local training set introduces the additional deterministic remainder $\delta_{\mathrm{rep},c}$. The next lemma connects this remainder to the local optimization trajectory.

Lemma 6 (Momentum and group-rate optimizer-to-representation stability)**.** *Consider two coupled local runs with the same initialization and zero initial momentum buffers. Let g_j and g_j^0 denote the exact implemented update directions, including the common clipping and weight-decay operations. For momentum $m \in [0,1)$, base learning rate $\eta_{\mathrm{lr},j}$, and a diagonal group-rate matrix P_j, the coupled recursions are*

$$\begin{aligned}v_{j+1} &= mv_j + g_j(\theta_j), \qquad \theta_{j+1} = \theta_j - \eta_{\mathrm{lr},j} P_j v_{j+1}, \\ v_{j+1}^0 &= mv_j^0 + g_j^0(\theta_j^0), \qquad \theta_{j+1}^0 = \theta_j^0 - \eta_{\mathrm{lr},j} P_j v_{j+1}^0, \qquad j = 0, \dots, J-1.\end{aligned}$$

Suppose every g_j is L_g-Lipschitz, $\sup_\theta \|g_j(\theta) - g_j^0(\theta)\|_2 \le b_j$, and $\|P_j\|_2 \le p_j$. Define

$$A_j = \begin{pmatrix} 1 + \eta_{\mathrm{lr},j} p_j L_g & \eta_{\mathrm{lr},j} p_j m \\ L_g & m \end{pmatrix}, \quad u_j = \begin{pmatrix} \eta_{\mathrm{lr},j} p_j b_j \\ b_j \end{pmatrix}. \tag{38}$$

If the normalized feature map is uniformly L_f-Lipschitz in θ, then

$$\begin{aligned}\varepsilon_{\mathrm{rep}} &\le L_f \|\theta_J - \theta_J^0\|_2 \\ &\le L_f \begin{pmatrix} 1 & 0 \end{pmatrix} \sum_{j=0}^{J-1} \left(\prod_{\ell=j+1}^{J-1} A_\ell \right) u_j.\end{aligned} \tag{39}$$

Here an empty product is the 2×2 identity and products are ordered with the larger index on the left. The experiment is the special case $m = 0.9$, constant base rate $\eta_{\mathrm{lr},j} \equiv 0.01$, and $P_j = P = \mathrm{diag}(1$ on LoRA/head$, 0.1$ on LayerNorm$)$, so $p_j = 1$. For an unchanged mini-batch, $b_j = 0$. If a batch-average gradient of size B changes in at most r_j examples and each per-example gradient has norm at most G, then one may take $b_j \le 2r_j G/B$.

Proof. Let $d_j^\theta = \|\theta_j - \theta_j^0\|_2$ and $d_j^v = \|v_j - v_j^0\|_2$. Adding and subtracting $g_j(\theta_j^0)$ gives

$$d_{j+1}^v \le L_g d_j^\theta + m d_j^v + b_j.$$

The parameter update and $\|P_j\|_2 \le p_j$ then imply

$$d_{j+1}^\theta \le (1 + \eta_{\mathrm{lr},j} p_j L_g) d_j^\theta + \eta_{\mathrm{lr},j} p_j m d_j^v + \eta_{\mathrm{lr},j} p_j b_j.$$

Therefore

$$\begin{pmatrix} d_{j+1}^\theta \\ d_{j+1}^v \end{pmatrix} \preceq A_j \begin{pmatrix} d_j^\theta \\ d_j^v \end{pmatrix} + u_j,$$

where $\preceq$ is component-wise. The initial vector is zero, so unrolling the recursion proves the second line of Eq. (39). Uniform feature Lipschitzness gives the first line. For the batch statement, each replaced example changes the gradient sum by at most $2G$, and division by B completes the bound.

C.2 A NORMALIZATION EXPANSION ON THE UNIT SPHERE

The proofs require an explicit expansion of the map

$$N(z) = \frac{z}{\|z\|_2}$$

around a unit vector. This is where the normalization-induced error terms in the main theorem come from.

Lemma 7 (Expansion of the normalization map)**.** *Let $\mu \in \mathbb{S}^{d-1}$. Let $\xi \in \mathbb{R}^d$ satisfy $\|\xi\|_2 \le r < 1/2$. Write*

$$a = \langle \mu, \xi \rangle, \qquad b = \Pi_\mu \xi, \qquad \Pi_\mu = I - \mu\mu^\top.$$

Then $b \perp \mu$, $\xi = a\mu + b$, and

$$\mathrm{Norm}(\mu + \xi) = \mu + b - ab - \frac{1}{2}\|b\|_2^2 \mu + R(\xi), \tag{40}$$

where $\|R(\xi)\|_2 \le C\|\xi\|_2^3$ for an absolute constant C. Consequently, if ξ is tangent to the sphere at μ, namely $\langle \mu, \xi \rangle = 0$, then

$$\begin{aligned} \mathrm{Norm}(\mu + \xi) - \mu = \xi - \frac{1}{2}\|\xi\|_2^2 \mu \\ + O(\|\xi\|_2^3). \end{aligned} \tag{41}$$

Proof. Since $\xi = a\mu + b$ and $b \perp \mu$,

$$\|\mu + \xi\|_2^2 = \|(1+a)\mu + b\|_2^2 = (1+a)^2 + \|b\|_2^2. \tag{42}$$

Let

$$u = 2a + a^2 + \|b\|_2^2.$$

Then $\|\mu + \xi\|_2 = (1+u)^{1/2}$. For $\|\xi\|_2 \le r < 1/2$, the Taylor expansion of $(1+u)^{-1/2}$ gives

$$\frac{1}{\|\mu + \xi\|_2} = 1 - \frac{1}{2}u + \frac{3}{8}u^2 + O(|u|^3). \tag{43}$$

Keeping terms up to second order in ξ, and using $u = 2a + a^2 + \|b\|_2^2$, gives

$$\begin{aligned} 1 - \frac{1}{2}u + \frac{3}{8}u^2 &= 1 - a - \frac{1}{2}a^2 - \frac{1}{2}\|b\|_2^2 + \frac{3}{2}a^2 + O(\|\xi\|_2^3) \\ &= 1 - a + a^2 - \frac{1}{2}\|b\|_2^2 + O(\|\xi\|_2^3). \end{aligned} \tag{44}$$

Therefore

$$\begin{aligned} \mathrm{Norm}(\mu + \xi) &= ((1+a)\mu + b) \\ &\quad \cdot \left(1 - a + a^2 - \frac{1}{2}\|b\|_2^2\right) \\ &\quad + O(\|\xi\|_2^3) \\ &= \mu + b - ab - \frac{1}{2}\|b\|_2^2 \mu \\ &\quad + O(\|\xi\|_2^3). \end{aligned} \tag{45}$$

This proves Eq. (40). Setting $a = 0$ gives Eq. (41).

Assumption 8 (Local prototype perturbation)**.** *For each valid pair (k, c), the empirical prototype can be written as*

$$p_{k,c} = \mathrm{Norm}(\mu_{k,c} + \xi_{k,c}). \tag{46}$$

There exists a fixed constant $r_0 < 1/2$, independent of (k, c), such that

$$\|\xi_{k,c}\|_2 \le r_0 \qquad \textit{almost surely}. \tag{47}$$

$\mathbb{E}[\xi_{k,c}] = 0$. *For fixed constants* $C_{\mathrm{mom}}, C_{\mathrm{rt}} > 0$, *the tangent component* $\Pi_{\mu_{k,c}}\xi_{k,c}$ *satisfies*

$$\begin{aligned} \mathbb{E}\|\Pi_{\mu_{k,c}}\xi_{k,c}\|_2^2 &= \frac{\sigma_{k,c}^2}{n_{k,c}}, \\ \mathbb{E}\|\xi_{k,c}\|_2^3 &\le C_{\mathrm{mom}} \left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}. \end{aligned} \tag{48}$$

In addition, the radial–tangent third-order interaction is controlled by

$$\left\|\mathbb{E}\left[\langle \mu_{k,c}, \xi_{k,c}\rangle \Pi_{\mu_{k,c}}\xi_{k,c}\right]\right\|_2 \le C_{\mathrm{rt}} \left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}. \tag{49}$$

This condition holds with a zero left-hand side, for example, when the joint law of $a = \langle \mu_{k,c}, \xi_{k,c}\rangle$ *and* $b = \Pi_{\mu_{k,c}}\xi_{k,c}$ *is invariant under* $(a, b) \mapsto (a, -b)$.

For distinct valid clients $i \neq j$, *the prototype perturbations are conditionally independent given their population directions and local supports, so that*

$$\xi_{i,c} \perp\!\!\!\perp \xi_{j,c} \mid (\mu_{i,c}, \mu_{j,c}, n_{i,c}, n_{j,c}). \tag{50}$$

In particular, with $\zeta_{k,c} = p_{k,c} - \mu_{k,c}$,

$$\mathbb{E}\langle \zeta_{i,c}, \zeta_{j,c}\rangle = \langle \mathbb{E}\zeta_{i,c}, \mathbb{E}\zeta_{j,c}\rangle. \tag{51}$$

In the fixed-map reference experiment, the two split prototypes use the random construction in Lemma 4, with disjoint, near-balanced subsets of sizes $m_{k,c}^{(a)}$ *and* $m_{k,c}^{(b)}$, *where* $m_{k,c}^{(a)} + m_{k,c}^{(b)} = n_{k,c}$ *and* $|m_{k,c}^{(a)} - m_{k,c}^{(b)}| \le 1$. *They are generated by independent perturbations* $\xi_{k,c}^{(a)}$ *and* $\xi_{k,c}^{(b)}$, *with*

$$\begin{aligned} p_{k,c}^{(a)} &= \mathrm{Norm}(\mu_{k,c} + \xi_{k,c}^{(a)}), \\ p_{k,c}^{(b)} &= \mathrm{Norm}(\mu_{k,c} + \xi_{k,c}^{(b)}), \end{aligned} \tag{52}$$

and

$$\begin{aligned} \mathbb{E}\|\Pi_{\mu_{k,c}}\xi_{k,c}^{(a)}\|_2^2 &= \frac{\sigma_{k,c}^2}{m_{k,c}^{(a)}}, \\ \mathbb{E}\|\Pi_{\mu_{k,c}}\xi_{k,c}^{(b)}\|_2^2 &= \frac{\sigma_{k,c}^2}{m_{k,c}^{(b)}}. \end{aligned} \tag{53}$$

The split perturbations also satisfy $\|\xi_{k,c}^{(a)}\|_2 \le r_0$ *and* $\|\xi_{k,c}^{(b)}\|_2 \le r_0$ *almost surely. The third-moment and radial–tangent interaction bounds hold after replacing* $\sigma_{k,c}^2/n_{k,c}$ *by the corresponding* $\sigma_{k,c}^2/m_{k,c}^{(a)}$ *or* $\sigma_{k,c}^2/m_{k,c}^{(b)}$.

The almost-sure radius provides a local-regime condition under which the normalization expansion is controlled uniformly.

C.3 Split-half noise scale

The client reports

$$\begin{aligned} \widehat{\nu}_{k,c} &= \chi_{k,c}\left(1 - \langle p_{k,c}^{(a)}, p_{k,c}^{(b)}\rangle\right), \\ \chi_{k,c} &= \frac{2m_{k,c}^{(a)}m_{k,c}^{(b)}}{n_{k,c}^2}. \end{aligned} \tag{54}$$

The full-prototype angular noise is

$$\nu_{k,c} = \mathbb{E}\left[1 - \langle p_{k,c}, \mu_{k,c}\rangle\right]. \tag{55}$$

Lemma 9 (Full-prototype angular noise)**.** *Under Assumption 8,*

$$\nu_{k,c} = \frac{1}{2}\frac{\sigma_{k,c}^2}{n_{k,c}} + O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right]. \tag{56}$$

Proof. For readability write $\mu = \mu_{k,c}$, $p = p_{k,c}$, and $\xi = \xi_{k,c}$. By Eq. (33),

$$1 - \langle p, \mu\rangle = \frac{1}{2}\|p - \mu\|_2^2. \tag{57}$$

By Lemma 7,

$$p - \mu = \Pi_\mu \xi - \langle \mu, \xi\rangle \Pi_\mu \xi - \frac{1}{2}\|\Pi_\mu \xi\|_2^2 \mu + O(\|\xi\|_2^3). \tag{58}$$

The first term is first order in ξ, while all other displayed terms are second order or higher. Squaring gives

$$\|p - \mu\|_2^2 = \|\Pi_\mu \xi\|_2^2 + O(\|\xi\|_2^3). \tag{59}$$

Taking expectations and using Assumption 8,

$$\begin{aligned}\mathbb{E}\|p - \mu\|_2^2 &= \mathbb{E}\|\Pi_\mu \xi\|_2^2 + O(\mathbb{E}\|\xi\|_2^3) \\ &= \frac{\sigma_{k,c}^2}{n_{k,c}} + O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right].\end{aligned} \tag{60}$$

Substituting Eq. (60) into Eq. (57) proves Eq. (56).

Lemma 10 (Equivalence of angular and tangent-noise scales)**.** *Under Assumption 8, constants $c_-, c_+ > 0$, depending only on the fixed radius r_0, exist such that*

$$c_- \frac{\sigma_{k,c}^2}{n_{k,c}} \le \nu_{k,c} \le c_+ \frac{\sigma_{k,c}^2}{n_{k,c}}. \tag{61}$$

Consequently,

$$O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right] = O(\nu_{k,c}^{3/2}),$$

with the convention that both sides vanish when $\sigma_{k,c}^2/n_{k,c} = 0$.

Proof. Suppress the indices and write

$$a = \langle \mu, \xi\rangle, \qquad b = \Pi_\mu \xi, \qquad s = \sqrt{(1+a)^2 + \|b\|_2^2}.$$

Because $\|\xi\|_2 \le r_0 < 1/2$, we have

$$1 - r_0 \le 1 + a \le 1 + r_0, \qquad 1 - r_0 \le s \le 1 + r_0.$$

Direct rationalization gives

$$\begin{aligned}1 - \langle \mathrm{Norm}(\mu + \xi), \mu\rangle &= 1 - \frac{1+a}{s} \\ &= \frac{\|b\|_2^2}{s\{s + (1+a)\}}.\end{aligned} \tag{62}$$

The denominator in Eq. (62) is bounded below by $2(1 - r_0)^2$ and above by $2(1 + r_0)^2$. Hence

$$\frac{\|b\|_2^2}{2(1+r_0)^2} \le 1 - \langle \mathrm{Norm}(\mu + \xi), \mu\rangle \le \frac{\|b\|_2^2}{2(1-r_0)^2}.$$

Taking expectations and using $\mathbb{E}\|b\|_2^2 = \sigma_{k,c}^2/n_{k,c}$ proves Eq. (61), with

$$c_- = \frac{1}{2(1+r_0)^2}, \qquad c_+ = \frac{1}{2(1-r_0)^2}.$$

The final order relation follows immediately from the two-sided bound.

Lemma 11 (Mean shrinkage caused by normalization)**.** *Under Assumption 8,*

$$\mathbb{E}[p_{k,c} - \mu_{k,c}] = -\nu_{k,c}\mu_{k,c} + O(\nu_{k,c}^{3/2}). \tag{63}$$

Proof. Using the notation of Lemma 9, the normalization expansion gives

$$p - \mu = \Pi_\mu \xi - \langle \mu, \xi \rangle \Pi_\mu \xi - \frac{1}{2}\|\Pi_\mu \xi\|_2^2 \mu + O(\|\xi\|_2^3). \tag{64}$$

Taking expectations term by term gives the three contributions below. First, $\mathbb{E}\Pi_\mu \xi = \Pi_\mu \mathbb{E}\xi = 0$, because the perturbation is centered. Second, Assumption 8 gives

$$\|\mathbb{E}\left[\langle \mu, \xi \rangle \Pi_\mu \xi\right]\|_2 = O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right]. \tag{65}$$

Third, the remainder is controlled by

$$\mathbb{E}\|O(\|\xi\|_2^3)\|_2 = O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right]. \tag{66}$$

Hence

$$\mathbb{E}[p - \mu] = -\frac{1}{2}\mathbb{E}\|\Pi_\mu \xi\|_2^2 \mu + O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right]. \tag{67}$$

By Lemma 9,

$$\nu_{k,c} = \frac{1}{2}\mathbb{E}\|\Pi_\mu \xi\|_2^2 + O\left[\left(\frac{\sigma_{k,c}^2}{n_{k,c}}\right)^{3/2}\right]. \tag{68}$$

Lemma 10 converts every remaining $O((\sigma_{k,c}^2/n_{k,c})^{3/2})$ term into $O(\nu_{k,c}^{3/2})$. Replacing the leading quadratic term by $\nu_{k,c}$ therefore yields Eq. (63).

Lemma 12 (Scale of the generalized split-half estimate)**.** *Under Assumption 8,*

$$\mathbb{E}[\widehat{\nu}_{k,c}] = 2\nu_{k,c} + O(\nu_{k,c}^{3/2}). \tag{69}$$

Proof. Write $m_a = m_{k,c}^{(a)}$, $m_b = m_{k,c}^{(b)}$, $n = n_{k,c} = m_a + m_b$, and $\chi = 2m_a m_b/n^2$. Let $\nu^{(a)}$ and $\nu^{(b)}$ be the angular noises of the two subset prototypes. Lemma 9 applied at the two subset sizes gives

$$\begin{aligned} \nu^{(a)} &= \frac{1}{2}\frac{\sigma_{k,c}^2}{m_a} + O\left[\left(\frac{\sigma_{k,c}^2}{m_a}\right)^{3/2}\right], \\ \nu^{(b)} &= \frac{1}{2}\frac{\sigma_{k,c}^2}{m_b} + O\left[\left(\frac{\sigma_{k,c}^2}{m_b}\right)^{3/2}\right]. \end{aligned} \tag{70}$$

Using independence of the two disjoint subset perturbations and the shrinkage expansion from Lemma 11,

$$\mathbb{E}\Big[1 - \langle p_{k,c}^{(a)}, p_{k,c}^{(b)} \rangle\Big] = \nu^{(a)} + \nu^{(b)} + O(\nu_{k,c}^{3/2}), \tag{71}$$

where near balance implies $m_a, m_b = \Theta(n)$, so all subset remainders have the full-prototype order $O(\nu_{k,c}^{3/2})$. Multiplying the leading term in Eq. (71) by χ yields

$$\begin{aligned} \chi \frac{\sigma_{k,c}^2}{2}\left(\frac{1}{m_a} + \frac{1}{m_b}\right) &= \frac{2m_a m_b}{n^2}\frac{\sigma_{k,c}^2}{2}\frac{n}{m_a m_b} \\ &= \frac{\sigma_{k,c}^2}{n}. \end{aligned} \tag{72}$$

By Lemma 9, $2\nu_{k,c} = \sigma_{k,c}^2/n + O(\nu_{k,c}^{3/2})$. Combining Eqs. (54)–(72) proves Eq. (69). When $m_a = m_b = n/2$, $\chi = 1/2$ and the estimator reduces to the standard equal-half form.

C.4 DISPERSION DECOMPOSITION

For $M_c > 1$, define

$$Z_c = 1 - \sum_{k \in \mathcal{K}_c^t} w_{k,c}^2 = \sum_{i \neq j} w_{i,c} w_{j,c} > 0. \tag{73}$$

The population cross-client representation dispersion is

$$D_c^\star = \frac{\sum_{i \neq j} w_{i,c} w_{j,c} \left(1 - \langle \mu_{i,c}, \mu_{j,c} \rangle\right)}{Z_c}. \tag{74}$$

This is equivalent to the main-text form

$$D_c^\star = 1 - \frac{\sum_{i \neq j} w_{i,c} w_{j,c} \langle \mu_{i,c}, \mu_{j,c} \rangle}{Z_c},$$

because $\sum_{i \neq j} w_{i,c} w_{j,c} = Z_c$. The empirical raw dispersion is

$$\widetilde{d}_c^t = \frac{\sum_{i \neq j} w_{i,c} w_{j,c} \left(1 - \langle p_{i,c}, p_{j,c} \rangle\right)}{Z_c}. \tag{75}$$

The pairwise-matched noise floor is

$$\widetilde{\nu}_c^t = \frac{\sum_{i \neq j} w_{i,c} w_{j,c} \frac{1}{2} \left(\widehat{\nu}_{i,c} + \widehat{\nu}_{j,c}\right)}{Z_c}. \tag{76}$$

Equivalently,

$$\widetilde{\nu}_c^t = \frac{\sum_i w_{i,c}(1 - w_{i,c}) \widehat{\nu}_{i,c}}{Z_c}, \tag{77}$$

because

$$\sum_{j \neq i} w_{i,c} w_{j,c} = w_{i,c} \sum_{j \neq i} w_{j,c} = w_{i,c}(1 - w_{i,c}).$$

Lemma 13 (Pairwise empirical distance decomposition)**.** *For any two distinct valid clients $i \neq j$,*

$$\begin{aligned} &\mathbb{E}\left[1 - \langle p_{i,c}, p_{j,c} \rangle\right] \\ &\quad = 1 - \langle \mu_{i,c}, \mu_{j,c} \rangle + \nu_{i,c} + \nu_{j,c} \\ &\qquad + O((\nu_{i,c} + \nu_{j,c})(1 - \langle \mu_{i,c}, \mu_{j,c} \rangle)) + O(\nu_{c,\max}^{3/2}), \end{aligned} \tag{78}$$

where

$$\nu_{c,\max} = \max_{k \in \mathcal{K}_c^t} \nu_{k,c}.$$

Proof. Suppress c for readability and write

$$\zeta_i = p_i - \mu_i, \qquad \zeta_j = p_j - \mu_j.$$

Then

$$\begin{aligned} \langle p_i, p_j \rangle &= \langle \mu_i + \zeta_i, \mu_j + \zeta_j \rangle \\ &= \langle \mu_i, \mu_j \rangle + \langle \zeta_i, \mu_j \rangle + \langle \mu_i, \zeta_j \rangle + \langle \zeta_i, \zeta_j \rangle. \end{aligned} \tag{79}$$

Therefore

$$\begin{aligned} &1 - \langle p_i, p_j \rangle \\ &\quad = 1 - \langle \mu_i, \mu_j \rangle - \langle \zeta_i, \mu_j \rangle \\ &\qquad - \langle \mu_i, \zeta_j \rangle - \langle \zeta_i, \zeta_j \rangle. \end{aligned} \tag{80}$$

By Lemma 11,

$$\begin{aligned} \mathbb{E}\zeta_i &= -\nu_i \mu_i + O(\nu_i^{3/2}), \\ \mathbb{E}\zeta_j &= -\nu_j \mu_j + O(\nu_j^{3/2}). \end{aligned}$$

Thus

$$-\mathbb{E}\langle \zeta_i, \mu_j \rangle = \nu_i \langle \mu_i, \mu_j \rangle + O(\nu_{c,\max}^{3/2}), \tag{81}$$

and

$$-\mathbb{E}\langle \mu_i, \zeta_j \rangle = \nu_j \langle \mu_i, \mu_j \rangle + O(\nu_{c,\max}^{3/2}). \tag{82}$$

By the cross-client conditional independence in Eq. (50),

$$\mathbb{E}\langle \zeta_i, \zeta_j \rangle = \langle \mathbb{E}\zeta_i, \mathbb{E}\zeta_j \rangle = \nu_i \nu_j \langle \mu_i, \mu_j \rangle + O(\nu_{c,\max}^{5/2}),$$

so

$$-\mathbb{E}\langle \zeta_i, \zeta_j \rangle = O(\nu_{c,\max}^2). \tag{83}$$

Because $0 \le \nu_{c,\max} \le 2$, the term $O(\nu_{c,\max}^2)$ is absorbed into $O(\nu_{c,\max}^{3/2})$. Taking expectations in Eq. (80) and substituting Eqs. (81)–(83) gives

$$\begin{aligned} &\mathbb{E}[1 - \langle p_i, p_j \rangle] \\ &\quad = 1 - \langle \mu_i, \mu_j \rangle + (\nu_i + \nu_j)\langle \mu_i, \mu_j \rangle + O(\nu_{c,\max}^{3/2}). \end{aligned} \tag{84}$$

Since

$$\langle \mu_i, \mu_j \rangle = 1 - (1 - \langle \mu_i, \mu_j \rangle),$$

we have

$$(\nu_i + \nu_j)\langle \mu_i, \mu_j \rangle = \nu_i + \nu_j - (\nu_i + \nu_j)(1 - \langle \mu_i, \mu_j \rangle). \tag{85}$$

Substituting Eq. (85) into Eq. (84) proves the claim.

Theorem 14 (Noise-calibrated dispersion)**.** *Under Assumption 8, for $M_c > 1$,*

$$\mathbb{E}[\widetilde{d}_c^t - \widetilde{\nu}_c^t] = D_c^\star + O(\nu_{c,\max} D_c^\star) + O(\nu_{c,\max}^{3/2}). \tag{86}$$

For a data-dependent fitted encoder satisfying Eq. (37)*, the implemented quantities instead satisfy*

$$\mathbb{E}[\widetilde{d}_c^t - \widetilde{\nu}_c^t] = D_c^\star + O(\nu_{c,\max} D_c^\star) + O(\nu_{c,\max}^{3/2}) + O(\delta_{\mathrm{rep},c}), \tag{87}$$

where $\delta_{\mathrm{rep},c} = 5 \max_{k \in \mathcal{K}_c^t} a_{\mathrm{rep},k,c}$.

Proof. Define

$$\begin{aligned} N_c &= \frac{\sum_{i \ne j} w_{i,c} w_{j,c} (\nu_{i,c} + \nu_{j,c})}{Z_c}, \\ H_c &= \frac{\sum_{i \ne j} w_{i,c} w_{j,c} (\nu_{i,c} + \nu_{j,c}) \delta_{ij,c}}{Z_c}, \end{aligned} \tag{88}$$

where $\delta_{ij,c} = 1 - \langle \mu_{i,c}, \mu_{j,c} \rangle$. Taking the weighted average of Lemma 13 with coefficients $w_{i,c} w_{j,c} / Z_c$, we get

$$\mathbb{E}[\widetilde{d}_c^t] = D_c^\star + N_c + O(H_c) + O(\nu_{c,\max}^{3/2}). \tag{89}$$

Here $D_c^\star$ is exactly the population dispersion from Eq. (74). Since $\nu_{i,c} + \nu_{j,c} \le 2\nu_{c,\max}$, the definition of H_c gives

$$O(H_c) = O(\nu_{c,\max} D_c^\star).$$

It remains to match N_c with the expected noise floor. By Lemma 12,

$$\mathbb{E}[\widehat{\nu}_{i,c}] = 2\nu_{i,c} + O(\nu_{c,\max}^{3/2}), \qquad \mathbb{E}[\widehat{\nu}_{j,c}] = 2\nu_{j,c} + O(\nu_{c,\max}^{3/2}).$$

Hence

$$\mathbb{E}\left[\frac{1}{2}(\widehat{\nu}_{i,c} + \widehat{\nu}_{j,c})\right] = \nu_{i,c} + \nu_{j,c} + O(\nu_{c,\max}^{3/2}). \tag{90}$$

Averaging Eq. (90) with the same normalized weights $w_{i,c} w_{j,c} / Z_c$ gives

$$\mathbb{E}[\widetilde{\nu}_c^t] = N_c + O(\nu_{c,\max}^{3/2}). \tag{91}$$

Subtracting Eq. (91) from Eq. (89) yields Eq. (86). For the implemented statistic, Eq. (37) transfers this fixed-map result to the fitted-and-reused map with an additive remainder bounded by $\delta_{\mathrm{rep},c}$. This proves Eq. (87).

Score interpretation. Theorem 14 concerns the instantaneous untruncated quantity $\widetilde{d}_c^t - \widetilde{\nu}_c^t$. The implemented controller smooths $\widetilde{d}_c^t$ and $\widetilde{\nu}_c^t$ separately and then uses $e_c^t = [d_c^t - \nu_c^t]_+$. The positive part prevents negative excess when finite-sample fluctuation makes the empirical floor larger than the raw dispersion; it does not preserve exact unbiasedness. Equation (87) is the applicable statement when the encoder is fitted and summarized on the same local training set.

Lemma 15 (EMA lag with intermittent valid updates)**.** *Fix a class c and a wall-clock round t. Let*

$$\mathcal{S}_c(t) = \{t_0 < t_1 < \cdots < t_m\} \subseteq \{0, \dots, t\}$$

be the rounds up to t at which at least two valid prototypes are available for class c. Write $x_c^{t_j} = \widetilde{d}_c^{t_j} - \widetilde{\nu}_c^{t_j}$, and let $y_c^{[0]} = y_c^0$ be the stored difference before the first valid update. The stored EMA after the j-th valid update is

$$y_c^{[j+1]} = \mu_{\mathrm{ema}} y_c^{[j]} + (1 - \mu_{\mathrm{ema}}) x_c^{t_j}. \tag{92}$$

Between t_j and t_{j+1}, the stored value remains $y_c^{[j+1]}$. If

$$\mathbb{E}[x_c^{t_j}] = D_c^{\star,t_j} + \varepsilon_c^{t_j},$$

then, after the last valid update not exceeding t,

$$\begin{aligned} &\left|\mathbb{E}[y_c^{[m+1]}] - D_c^{\star,t}\right| \\ &\quad \le \mu_{\mathrm{ema}}^{m+1} \left|y_c^0 - D_c^{\star,t}\right| \\ &\qquad + (1 - \mu_{\mathrm{ema}}) \sum_{j=0}^{m} \mu_{\mathrm{ema}}^{m-j} \Delta_c^{t_j,t}, \end{aligned} \tag{93}$$

where

$$\Delta_c^{t_j,t} = |\varepsilon_c^{t_j}| + |D_c^{\star,t_j} - D_c^{\star,t}|.$$

If $\mathcal{S}_c(t) = \varnothing$, the state is unchanged and the left-hand side is exactly $|y_c^0 - D_c^{\star,t}|$.

Proof. Repeatedly applying Eq. (92) over the valid-update index j, rather than over every wall-clock round, gives

$$\begin{aligned} \mathbb{E}[y_c^{[m+1]}] &= \mu_{\mathrm{ema}}^{m+1} y_c^0 \\ &\quad + (1 - \mu_{\mathrm{ema}}) \sum_{j=0}^{m} \mu_{\mathrm{ema}}^{m-j} \left(D_c^{\star,t_j} + \varepsilon_c^{t_j}\right). \end{aligned} \tag{94}$$

Subtract $D_c^{\star,t}$. Since

$$(1 - \mu_{\mathrm{ema}}) \sum_{j=0}^{m} \mu_{\mathrm{ema}}^{m-j} = 1 - \mu_{\mathrm{ema}}^{m+1},$$

we obtain

$$\begin{aligned} &\mathbb{E}[y_c^{[m+1]}] - D_c^{\star,t} \\ &= \mu_{\mathrm{ema}}^{m+1} \left(y_c^0 - D_c^{\star,t}\right) \\ &\quad + (1 - \mu_{\mathrm{ema}}) \sum_{j=0}^{m} \mu_{\mathrm{ema}}^{m-j} \\ &\qquad \times \left(D_c^{\star,t_j} - D_c^{\star,t} + \varepsilon_c^{t_j}\right). \end{aligned} \tag{95}$$

Applying the triangle inequality proves Eq. (93). Invalid rounds perform no recursion, so retained state is represented exactly by the valid-update index set $\mathcal{S}_c(t)$.

C.5 Single- and multi-prototype approximation

For a fixed class c, let the population directions be $\{\mu_{k,c}\}_{k \in \mathcal{K}}$ with weights $\{w_{k,c}\}_{k \in \mathcal{K}}$. Define the weighted Euclidean mean

$$\bar{\mu}_c = \sum_{k \in \mathcal{K}} w_{k,c} \mu_{k,c}. \tag{96}$$

When the class is fixed, we suppress the subscript c and write μ_k, w_k, and $\bar{\mu}$ in the proofs below.

Lemma 16 (Weighted pairwise variance identity)**.** *For any vectors $u_1, \dots, u_M$ and weights $w_k \geq 0$ with $\sum_k w_k = 1$, if $\bar{u} = \sum_k w_k u_k$, then*

$$\sum_k w_k \|u_k - \bar{u}\|_2^2 = \frac{1}{2} \sum_{i,j} w_i w_j \|u_i - u_j\|_2^2. \tag{97}$$

Proof. Expanding the left-hand side gives

$$\begin{aligned}
&\sum_k w_k \|u_k - \bar{u}\|_2^2 \\
&\quad = \sum_k w_k \left(\|u_k\|_2^2 - 2\langle u_k, \bar{u} \rangle + \|\bar{u}\|_2^2 \right) \\
&\quad = \sum_k w_k \|u_k\|_2^2 - 2 \left\langle \sum_k w_k u_k, \bar{u} \right\rangle \\
&\qquad + \|\bar{u}\|_2^2 \sum_k w_k \\
&\quad = \sum_k w_k \|u_k\|_2^2 - 2\|\bar{u}\|_2^2 + \|\bar{u}\|_2^2 \\
&\quad = \sum_k w_k \|u_k\|_2^2 - \|\bar{u}\|_2^2.
\end{aligned} \tag{98}$$

Expanding the right-hand side gives

$$\begin{aligned}
&\frac{1}{2} \sum_{i,j} w_i w_j \|u_i - u_j\|_2^2 \\
&\quad = \frac{1}{2} \sum_{i,j} w_i w_j \big(\|u_i\|_2^2 + \|u_j\|_2^2 - 2\langle u_i, u_j \rangle \big) \\
&\quad = \frac{1}{2} \sum_i w_i \|u_i\|_2^2 \sum_j w_j \\
&\qquad + \frac{1}{2} \sum_j w_j \|u_j\|_2^2 \sum_i w_i \\
&\qquad - \sum_{i,j} w_i w_j \langle u_i, u_j \rangle \\
&\quad = \sum_i w_i \|u_i\|_2^2 - \left\langle \sum_i w_i u_i, \sum_j w_j u_j \right\rangle \\
&\quad = \sum_i w_i \|u_i\|_2^2 - \|\bar{u}\|_2^2.
\end{aligned} \tag{99}$$

The two expressions are identical.

Theorem 17 (Weighted Euclidean consensus identity)**.** *For unit population directions,*

$$\begin{aligned}
E_{\mathrm{euc}}(c) &= \sum_k w_{k,c} \|\mu_{k,c} - \bar{\mu}_c\|_2^2 \\
&= \sum_{i \neq j} w_{i,c} w_{j,c} \delta_{ij,c}.
\end{aligned} \tag{100}$$

Equivalently,

$$E_{\mathrm{euc}}(c) = Z_c D_c^\star. \tag{101}$$

Proof. Apply Lemma 16 with $u_k = \mu_{k,c}$. Since each $\mu_{k,c}$ is unit-normalized,

$$\frac{1}{2} \|\mu_{i,c} - \mu_{j,c}\|_2^2 = 1 - \langle \mu_{i,c}, \mu_{j,c} \rangle.$$

Thus

$$\sum_k w_{k,c}\|\mu_{k,c} - \bar{\mu}_c\|_2^2 = \sum_{i,j} w_{i,c}w_{j,c}\left(1 - \langle \mu_{i,c}, \mu_{j,c}\rangle\right).$$

The terms with $i = j$ vanish because $1 - \langle \mu_{i,c}, \mu_{i,c}\rangle = 0$. This gives Eq. (100). Dividing and multiplying by $Z_c = \sum_{i\neq j} w_{i,c}w_{j,c}$ gives Eq. (101).

Lemma 18 (Relation to a unit-norm single prototype)**.** *Let*

$$E_{\text{unit}}(c) = \min_{\|b\|_2=1} \sum_k w_{k,c}\|\mu_{k,c} - b\|_2^2.$$

If $\bar{\mu}_c \neq 0$, the minimizer is $b^\star = \bar{\mu}_c/\|\bar{\mu}_c\|_2$; if $\bar{\mu}_c = 0$, every unit vector is optimal. In either case,

$$E_{\text{unit}}(c) = 2(1 - \|\bar{\mu}_c\|_2). \tag{102}$$

The Euclidean mean error is

$$E_{\text{euc}}(c) = 1 - \|\bar{\mu}_c\|_2^2 = (1 + \|\bar{\mu}_c\|_2)(1 - \|\bar{\mu}_c\|_2), \tag{103}$$

so

$$E_{\text{euc}}(c) \leq E_{\text{unit}}(c) \leq 2E_{\text{euc}}(c). \tag{104}$$

Proof. For any unit b,

$$\begin{aligned}\sum_k w_k\|\mu_k - b\|_2^2 &= \sum_k w_k\left(\|\mu_k\|_2^2 + \|b\|_2^2 - 2\langle \mu_k, b\rangle\right)\\ &= 2 - 2\left\langle \sum_k w_k\mu_k, b\right\rangle\\ &= 2 - 2\langle \bar{\mu}, b\rangle.\end{aligned} \tag{105}$$

Minimizing this expression is equivalent to maximizing $\langle \bar{\mu}, b\rangle$ over $\|b\|_2 = 1$. If $\bar{\mu} = 0$, the objective equals 2 for every unit b. Otherwise, the Cauchy–Schwarz inequality shows that the maximum is $\|\bar{\mu}\|_2$, attained at $b^\star = \bar{\mu}/\|\bar{\mu}\|_2$. This proves Eq. (102). For the Euclidean mean error,

$$E_{\text{euc}} = \sum_k w_k\|\mu_k\|_2^2 - \|\bar{\mu}\|_2^2 = 1 - \|\bar{\mu}\|_2^2.$$

Since $0 \leq \|\bar{\mu}\|_2 \leq 1$,

$$1 - \|\bar{\mu}\|_2 \leq 1 - \|\bar{\mu}\|_2^2 \leq 2(1 - \|\bar{\mu}\|_2),$$

which proves Eq. (104).

Theorem 19 (Multi-prototype approximation)**.** *Define the optimal weighted approximation error with $M \geq 1$ retained prototypes by*

$$E_M^\star(c) = \min_{\{b_m\}_{m=1}^M} \sum_k w_{k,c} \min_{1\leq m\leq M} \|\mu_{k,c} - b_m\|_2^2. \tag{106}$$

Then

$$E_M^\star(c) \leq E_1^\star(c). \tag{107}$$

For $M = 1$, the minimizer is the weighted Euclidean mean, so

$$E_1^\star(c) = E_{\text{euc}}(c) = Z_c D_c^\star. \tag{108}$$

Consequently,

$$0 \leq E_1^\star(c) - E_M^\star(c) \leq Z_c D_c^\star. \tag{109}$$

If all retained prototypes are constrained to the unit sphere and $E_{M,\text{unit}}^\star(c)$ denotes the corresponding optimum, then

$$0 \leq E_{1,\text{unit}}^\star(c) - E_{M,\text{unit}}^\star(c) \leq 2Z_c D_c^\star. \tag{110}$$

The unit-norm approximation error is also nonincreasing in M.

Proof. Let $b_1^\star$ be an optimizer of the one-prototype problem. For the M-prototype problem, choose

$$b_1 = b_1^\star, \qquad b_2 = \cdots = b_M = b_1^\star.$$

This choice is feasible for the M-prototype problem. For every k,

$$\min_{1\le m\le M} \|\mu_{k,c} - b_m\|_2^2 = \|\mu_{k,c} - b_1^\star\|_2^2.$$

Therefore the M-prototype objective can achieve the one-prototype error, and the optimum over M prototypes cannot be larger. This proves Eq. (107). Since $E_M^\star(c) \ge 0$,

$$0 \le E_1^\star(c) - E_M^\star(c) \le E_1^\star(c).$$

For $M = 1$, minimizing over an unconstrained center gives the weighted mean $\bar\mu_c$, and Theorem 17 therefore gives $E_1^\star(c) = E_{\mathrm{euc}}(c) = Z_c D_c^\star$. Under unit-sphere constraints, duplicating an optimizer for the smaller bank gives the same monotonicity argument. Lemma 18 gives $E_{1,\mathrm{unit}}^\star(c) \le 2E_{\mathrm{euc}}(c) = 2Z_c D_c^\star$, and nonnegativity then yields Eq. (110).

Corollary 20 (Strict Euclidean gain under separated modes)**.** *Suppose the valid clients for class c can be partitioned into Q nonempty groups*

$$\mathcal{G}_1, \dots, \mathcal{G}_Q.$$

Let

$$W_q = \sum_{k\in\mathcal{G}_q} w_{k,c}, \qquad \bar\mu_{q,c} = \frac{1}{W_q}\sum_{k\in\mathcal{G}_q} w_{k,c}\mu_{k,c}.$$

For $M \ge Q$,

$$E_M^\star(c) \le \sum_{q=1}^{Q}\sum_{k\in\mathcal{G}_q} w_{k,c}\|\mu_{k,c} - \bar\mu_{q,c}\|_2^2. \tag{111}$$

Expanding the same sum,

$$\begin{aligned} E_{\mathrm{euc}}(c) = &\sum_{q=1}^{Q}\sum_{k\in\mathcal{G}_q} w_{k,c}\|\mu_{k,c} - \bar\mu_{q,c}\|_2^2 \\ &+ \sum_{q=1}^{Q} W_q\|\bar\mu_{q,c} - \bar\mu_c\|_2^2. \end{aligned} \tag{112}$$

Thus, if at least one group mean differs from the global mean, the between-group term is positive and $E_M^\star(c) < E_{\mathrm{euc}}(c)$.

Proof. Set one retained prototype for each group, so that

$$b_q = \bar\mu_{q,c}, \qquad q = 1, \dots, Q.$$

Since $M \ge Q$, this is feasible. Assign every client $k \in \mathcal{G}_q$ to prototype b_q. The resulting objective value is exactly the right-hand side of Eq. (111); the optimum cannot be larger.

It remains to prove the decomposition. For $k \in \mathcal{G}_q$,

$$\mu_{k,c} - \bar\mu_c = (\mu_{k,c} - \bar\mu_{q,c}) + (\bar\mu_{q,c} - \bar\mu_c).$$

Squaring and summing with weights gives

$$\begin{aligned} &\sum_{k\in\mathcal{G}_q} w_{k,c}\|\mu_{k,c} - \bar\mu_c\|_2^2 \\ &= \sum_{k\in\mathcal{G}_q} w_{k,c}\|\mu_{k,c} - \bar\mu_{q,c}\|_2^2 + W_q\|\bar\mu_{q,c} - \bar\mu_c\|_2^2 \\ &\quad + 2\left\langle \sum_{k\in\mathcal{G}_q} w_{k,c}(\mu_{k,c} - \bar\mu_{q,c}), \bar\mu_{q,c} - \bar\mu_c \right\rangle. \end{aligned} \tag{113}$$

The cross term vanishes because

$$\sum_{k\in\mathcal{G}_q} w_{k,c}(\mu_{k,c} - \bar\mu_{q,c}) = \sum_{k\in\mathcal{G}_q} w_{k,c}\mu_{k,c} - W_q\bar\mu_{q,c} = 0.$$

Summing Eq. (113) over q proves Eq. (112). The between-group term is a sum of nonnegative squared norms and is positive whenever at least one $\bar\mu_{q,c} \ne \bar\mu_c$.

Proposition 21 (Transfer from reliability to support weights)**.** *For class c, with all sums below taken over the valid client set $\mathcal{K}_c^t$, let*

$$w_{k,c} = \frac{\sqrt{n_{k,c}}}{\sum_{j\in\mathcal{K}_c^t}\sqrt{n_{j,c}}}, \qquad r_{k,c} = \frac{n_{k,c}}{\sum_{j\in\mathcal{K}_c^t} n_{j,c}}.$$

For a weight vector ϖ on the valid clients, define

$$E^\star_{M,\varpi}(c) = \min_{\{b_m\}_{m=1}^M} \sum_k \varpi_{k,c} \min_m \|\mu_{k,c} - b_m\|_2^2. \tag{114}$$

The prototype centers may be restricted to the closed unit ball without increasing the objective. Consequently,

$$\left|E^\star_{M,w}(c) - E^\star_{M,r}(c)\right| \le 4\|w_c - r_c\|_1. \tag{115}$$

The same statement holds when all prototype centers are constrained to the unit sphere.

Proof. Every population direction $\mu_{k,c}$ lies in the closed unit ball. Projecting any candidate center onto that ball cannot increase its distance to any $\mu_{k,c}$, because Euclidean projection onto a closed convex set is nonexpansive and fixes every point in the set. Hence an optimum exists with $\|b_m\|_2 \le 1$, and therefore

$$0 \le \min_m \|\mu_{k,c} - b_m\|_2^2 \le 4.$$

For a fixed prototype set $B = \{b_m\}_{m=1}^M$, define

$$J_\varpi(B) = \sum_k \varpi_{k,c} \min_m \|\mu_{k,c} - b_m\|_2^2.$$

Then

$$\begin{aligned} |J_w(B) - J_r(B)| &\le \sum_k |w_{k,c} - r_{k,c}| \min_m \|\mu_{k,c} - b_m\|_2^2 \\ &\le 4\|w_c - r_c\|_1. \end{aligned} \tag{116}$$

Let $B_r^\star$ minimize J_r. Then

$$E^\star_{M,w} \le J_w(B_r^\star) \le E^\star_{M,r} + 4\|w_c - r_c\|_1.$$

Interchanging w and r gives the reverse inequality and proves Eq. (115). Under unit-sphere constraints, the squared distance is still at most 4, so the same proof applies.

Relation to the implemented bank. The population analysis above uses the reliability weights $w_{k,c}$, whereas the implemented bank uses the support weights $r_{k,c}$. Proposition 21 bounds the resulting change in the population objective. The implemented weighted spherical k-means further operates on empirical prototypes, so $E^\star_M(c)$ is a population reference for the score-controlled capacity rather than an exact characterization of the finite-iteration bank.

C.6 Score-to-operation monotonicity and alignment

The calibrated score is designed to track the relative strength of population dispersion against the matched finite-sample floor. The following results make the induced operation directions explicit. For nonnegative population dispersion D and matched noise level N, define the oracle score

$$s^\circ(D,N) = \begin{cases} D/(D+N), & D+N>0, \\ 0, & D+N=0. \end{cases} \tag{117}$$

Lemma 22 (Stability of the implemented stabilized score ratio)**.** *Let $h_\epsilon(e,n) = e/(e+n+\epsilon)$ for $e,n \ge 0$. If*

$$e+n+\epsilon \ge m_0 > 0, \qquad D+N \ge m_0 > 0,$$

then

$$|h_\epsilon(e,n) - s^\circ(D,N)| \le \frac{|e-D| + |n-N| + \epsilon}{m_0}. \tag{118}$$

Proof. On the nonnegative quadrant,

$$\left|\frac{\partial h_\epsilon}{\partial e}\right| = \frac{n+\epsilon}{(e+n+\epsilon)^2} \le \frac{1}{e+n+\epsilon}, \qquad \left|\frac{\partial h_\epsilon}{\partial n}\right| = \frac{e}{(e+n+\epsilon)^2} \le \frac{1}{e+n+\epsilon}.$$

Every point on the line segment joining (e, n) and (D, N) has coordinate sum plus ϵ at least m_0. The mean-value inequality bounds $|h_\epsilon(e, n) - h_\epsilon(D, N)|$ by the first two numerator terms in Eq. (118). For the remaining term,

$$\left|h_\epsilon(D, N) - \frac{D}{D+N}\right| = \frac{D\epsilon}{(D+N+\epsilon)(D+N)} \le \frac{\epsilon}{m_0}.$$

The triangle inequality proves the claim.

Proposition 23 (Direct bridge to the implemented positive-part/EMA score)**.** *At a valid update, let the stored EMA state be $d, \nu \ge 0$, let $e = [d-\nu]_+$, and let the implemented score be $s = h_\epsilon(e, \nu)$. For targets $D, N \ge 0$, suppose*

$$|(d-\nu) - D| \le r_D, \qquad |\nu - N| \le r_N,$$

and the two denominator conditions of Lemma 22 hold. Then

$$|s - s^\circ(D, N)| \le \frac{r_D + r_N + \epsilon}{m_0}. \tag{119}$$

If the two premise bounds hold on an event of probability at least $1-\delta$, Eq. (119) *holds on the same event. On a round without a valid update, the same statement applies to the retained state, with r_D, r_N including the EMA lag terms from Lemma 15.*

Proof. Projection onto the nonnegative half-line is nonexpansive and fixes $D \ge 0$, so

$$|e - D| = |[d-\nu]_+ - [D]_+| \le |(d-\nu) - D| \le r_D.$$

Apply Lemma 22 with $n = \nu$, and use $|\nu - N| \le r_N$. The event statement follows on the same event as the two premise bounds. Retention changes neither d nor ν, so only the target drift/lag component of their errors changes on an invalid round.

Proposition 24 (Monotone capacity and approximation)**.** *For fixed integers $M_c \ge 1$ and $M_{\max} \ge 1$, let*

$$M(s) = \min\left\{M_c,\, 1 + \left\lfloor s(M_{\max} - 1) + \tfrac{1}{2}\right\rfloor\right\}.$$

If $0 \le s_1 \le s_2 \le 1$, then

$$M(s_1) \le M(s_2), \qquad E^\star_{M(s_2)}(c) \le E^\star_{M(s_1)}(c). \tag{120}$$

Proof. The rounded affine expression inside the outer minimum is nondecreasing in s, so $M(s_1) \le M(s_2)$. The proof of Theorem 19 applies whenever one increases the number of available representatives, since duplicating an optimizer for the smaller bank is feasible for the larger bank. Hence the optimal approximation error cannot increase.

Proposition 25 (Direct displacement bound for classifier anchoring)**.** *If the anchoring operation is skipped because either normalized direction is invalid, its parameter displacement is zero. Otherwise, let $\widetilde{w}_c = \mathrm{Norm}(\widetilde{W}_c)$ be the unit-norm averaged classifier-row direction and $b_c = \bar{p}_c$ the unit-norm support-weighted bank consensus in Eq.* (15)*. For $\eta(s) = \eta_0(1-s)$, with $0 < \eta_0 < 1/2$, define the implemented post-anchoring direction*

$$a_c(s) = \mathrm{Norm}\left((1-\eta(s))\widetilde{w}_c + \eta(s) b_c\right).$$

Then

$$\|a_c(s) - \widetilde{w}_c\|_2^2 \le \frac{4\eta_0^2(1-s)^2}{(1-2\eta_0)^2}\|b_c - \widetilde{w}_c\|_2^2 \le \frac{16\eta_0^2(1-s)^2}{(1-2\eta_0)^2}. \tag{121}$$

Because Eq. (15) *restores the pre-anchoring row norm, the actual parameter displacement also satisfies*

$$\|\bar{W}_c - \widetilde{W}_c\|_2^2 = \|\widetilde{W}_c\|_2^2 \|a_c(s) - \widetilde{w}_c\|_2^2. \tag{122}$$

Thus increasing s decreases a pointwise upper bound on the displacement of the actual classifier row. The bound makes no assumption that $\widetilde{W}_c$ equals a population class direction or that the empirical bank consensus is an optimal population prototype.

Proof. For brevity, write $\eta = \eta(s)$ and set $u_c = (1-\eta)\widetilde{w}_c + \eta b_c$. Since both directions are unit vectors,

$$\|u_c - \widetilde{w}_c\|_2 = \eta\|b_c - \widetilde{w}_c\|_2, \qquad \|u_c\|_2 \geq 1 - 2\eta_0.$$

The normalization inequality used in Remark 5 gives

$$\|\operatorname{Norm}(u_c) - \widetilde{w}_c\|_2 \leq \frac{2\eta}{1-2\eta_0}\|b_c - \widetilde{w}_c\|_2.$$

Squaring, substituting $\eta = \eta_0(1-s)$, and using $\|b_c - \widetilde{w}_c\|_2 \leq 2$ proves Eq. (121).

Proposition 26 (A sufficient descent condition for the semantic gate)**.** *Fix a client and round. For every bank-covered class, abbreviate the non-score scarcity factor in Eq.* (1) *by $h^t_{k,c} \in [0,1]$, and write*

$$\vartheta_c(s_c) = h^t_{k,c}\{\gamma + (1-\gamma)s_c\}.$$

Let $R^{0,t}_{k,c}$ denote the corresponding ungated class component of the normalized semantic objective. The semantic-gradient contribution of the implemented objective is

$$S^t_k(\mathbf{s};\omega) = \lambda \sum_{c \in \mathcal{C}^t_{\mathcal{B}}} \vartheta_c(s_c)\nabla R^{0,t}_{k,c}(\omega).$$

At a point ω, write

$$A = \langle \nabla F_k(\omega), S^t_k(\mathbf{s};\omega)\rangle, \qquad B = \|S^t_k(\mathbf{s};\omega)\|_2^2.$$

For an L-smooth F_k, compare the corresponding full-gradient step

$$\omega^+(\mathbf{s}) = \omega - \eta_{\mathrm{lr}}\{\nabla F_k(\omega) + S^t_k(\mathbf{s};\omega)\}$$

with the plain-gradient update. If $A \geq 0$, $L\eta_{\mathrm{lr}} < 1$, and either $B = 0$ or

$$B \leq \frac{2(1-L\eta_{\mathrm{lr}})A}{L\eta_{\mathrm{lr}}}, \tag{123}$$

then the smoothness upper bound on $F_k(\omega^+(\mathbf{s}))$ is no larger than the corresponding upper bound for the plain-gradient update. Because B is the norm of the full sum, this condition includes every cross-class inner product and does not require class-wise semantic gradients to be orthogonal.

Proof. The L-smoothness bound for the gated step differs from that of the plain-gradient step by

$$-\eta_{\mathrm{lr}}(1-L\eta_{\mathrm{lr}})A + \frac{L\eta^2_{\mathrm{lr}}}{2}B. \tag{124}$$

This expression is nonpositive when $B = 0$. For $B > 0$, it is nonpositive under Eq. (123). Thus the gated full-gradient direction has a sufficient descent condition whenever the semantic and task gradients are nonnegatively aligned and the complete semantic contribution remains in the stated range.

Proposition 27 (Direction of the inference prior)**.** *Let the linear and prototype probability-error vectors be e_{lin} and $e_{\mathrm{proto}}(s)$. Define*

$$V_{\mathrm{lin}} = \mathbb{E}\|e_{\mathrm{lin}}\|_2^2, \quad V_{\mathrm{proto}}(s) = \mathbb{E}\|e_{\mathrm{proto}}(s)\|_2^2, \quad C(s) = \mathbb{E}\langle e_{\mathrm{lin}}, e_{\mathrm{proto}}(s)\rangle.$$

The mean-squared risk of a convex fusion with linear weight α is

$$\mathcal{E}_{\mathrm{fuse}}(\alpha;s) = \alpha^2 V_{\mathrm{lin}} + (1-\alpha)^2 V_{\mathrm{proto}}(s) + 2\alpha(1-\alpha)C(s). \tag{125}$$

If

$$D(s) = V_{\mathrm{lin}} + V_{\mathrm{proto}}(s) - 2C(s) > 0,$$

its unconstrained minimizer and convex-fusion minimizer are

$$\alpha_{\mathrm{unc}}(s) = \frac{V_{\mathrm{proto}}(s) - C(s)}{D(s)}, \qquad \alpha_{\mathrm{opt}}(s) = \Pi_{[0,1]}(\alpha_{\mathrm{unc}}(s)). \tag{126}$$

If V_{proto} and C are differentiable, then

$$\alpha'_{\mathrm{unc}}(s) = \frac{V'_{\mathrm{proto}}(s)(V_{\mathrm{lin}} - C(s)) + C'(s)(V_{\mathrm{proto}}(s) - V_{\mathrm{lin}})}{D(s)^2}. \tag{127}$$

Thus the risk-optimal linear weight is nonincreasing wherever the numerator in Eq. (127) *is nonpositive. Projection onto* $[0,1]$ *preserves that monotonicity. In particular,* $C(s) = 0$*, together with nonincreasing prototype risk, is one sufficient special case rather than an assumed identity. The implemented clamped affine prior in Eq.* (16) *is a bounded affine surrogate with the same monotone direction. Because the client supports are nonnegative,* $\bar{s}_k$ *is nondecreasing in every class score, so the implemented* $\widehat{\alpha}_k$ *is nonincreasing in every class score before and after clamping.*

Proof. Expanding the squared norm of $\alpha e_{\mathrm{lin}} + (1-\alpha)e_{\mathrm{proto}}$ gives Eq. (125). Its second derivative in α is $2D(s) > 0$. Setting its first derivative to zero gives α_{unc}; projection gives the constrained minimizer in Eq. (126). Differentiating that quotient and collecting terms yields Eq. (127). The weighted average $\bar{s}_k = \sum_c (n_{k,c}/n_k) s_c^\star$ has derivative $n_{k,c}/n_k \geq 0$ with respect to $s_c^\star$, and the affine map $1 - \beta \bar{s}_k$ is nonincreasing because $\beta \geq 0$. Clamping preserves monotonicity, while including $\alpha = 1$ in the candidate set ensures the optimality property in Proposition 37.

Scope of the operation results. Proposition 24 establishes monotone capacity and population approximation, while Proposition 25 controls the classifier-row displacement under the decreasing anchoring schedule. Propositions 26 and 27 give sufficient directional conditions for semantic gating and inference fusion, retaining the full cross-class and cross-error terms in their respective statements.

C.7 Score-gated objective and convergence

To distinguish the optimization learning rate from the anchoring coefficient η_c^t, we write the former as η_{lr} throughout this subsection. Let $\omega = (\Delta, \phi)$ denote the trainable LoRA and head/norm parameters. For a random local mini-batch $\mathcal{M} \sim \mathfrak{M}_k$, define

$$\mathcal{J}_k(\omega; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} \ell_{\mathrm{ce}}\left(z_{\mathrm{lin}}(x;\omega), y\right). \tag{128}$$

The plain local objective is

$$F_k(\omega) = \mathbb{E}_{\mathcal{M}\sim\mathfrak{M}_k}[\mathcal{J}_k(\omega;\mathcal{M})]. \tag{129}$$

At round t, the bank $\mathcal{B}^t$, covered-class set $\mathcal{C}_{\mathcal{B}}^t$, and gates $g_{k,c}^{\mathrm{eff},t}$ are fixed during local optimization. Define

$$\mathcal{R}_k^t(\omega;\mathcal{M}) = \frac{\sum_{(x,y)\in\mathcal{M}} \mathbf{1}[y \in \mathcal{C}_{\mathcal{B}}^t] g_{k,y}^{\mathrm{eff},t} \ell_{\mathrm{proto}}(x, y; \omega, \mathcal{B}^t)}{\sum_{(x,y)\in\mathcal{M}} \mathbf{1}[y \in \mathcal{C}_{\mathcal{B}}^t] + \epsilon}. \tag{130}$$

Its mini-batch expectation is

$$R_k^t(\omega) = \mathbb{E}_{\mathcal{M}\sim\mathfrak{M}_k}[\mathcal{R}_k^t(\omega;\mathcal{M})]. \tag{131}$$

Thus Eq. (130) is exactly the covered-sample normalization used by Eq. (3). When a mini-batch has no covered target, the numerator is zero and the semantic contribution is zero.

The regularized mini-batch and local objectives are

$$\begin{aligned} \widetilde{\mathcal{J}}_k^t(\omega;\mathcal{M}) &= \mathcal{J}_k(\omega;\mathcal{M}) + \lambda \mathcal{R}_k^t(\omega;\mathcal{M}), \\ \widetilde{F}_k^t(\omega) &= \mathbb{E}_{\mathcal{M}\sim\mathfrak{M}_k}[\widetilde{\mathcal{J}}_k^t(\omega;\mathcal{M})] \\ &= F_k(\omega) + \lambda R_k^t(\omega). \end{aligned} \tag{132}$$

Let $a_k = n_k / \sum_j n_j$. The plain and regularized global objectives are

$$\begin{aligned} F(\omega) &= \sum_k a_k F_k(\omega), \\ \widetilde{F}^t(\omega) &= \sum_k a_k \widetilde{F}_k^t(\omega) \\ &= F(\omega) + \lambda R^t(\omega), \end{aligned} \tag{133}$$

where $R^t(\omega) = \sum_k a_k R_k^t(\omega)$. Define $F^\star = \inf_\omega F(\omega) > -\infty$. The convergence analysis assumes full client participation, matching the main benchmark protocol. Hence the round-wise aggregation weights satisfy $a_k^t = a_k$. The partial-participation ablation is evaluated experimentally and is not included in the following bounds.

Assumption 28 (Optimization assumptions)**.** *The following hold for every round t.*

1. *Each F_k is L-smooth and each R_k^t is L_R-smooth. Hence each $\widetilde{F}_k^t$ and $\widetilde{F}^t$ is L_λ-smooth with $L_\lambda = L + \lambda L_R$.*
2. *Differentiation and mini-batch expectation may be interchanged, and the stochastic gradient of the implemented normalized mini-batch loss is conditionally unbiased, so that*
$$\mathbb{E}[g_{k,h}^t \mid \mathcal{F}_h^t] = \nabla \widetilde{F}_k^t(\omega_{k,h}^t),$$
where $\mathcal{F}_h^t$ contains the history before local step h.
3. *With*
$$\xi_{k,h}^t = g_{k,h}^t - \nabla \widetilde{F}_k^t(\omega_{k,h}^t),$$
the noises are martingale differences, are conditionally uncorrelated across clients, and satisfy
$$\mathbb{E}[\|\xi_{k,h}^t\|_2^2 \mid \mathcal{F}_h^t] \le \sigma_{\mathrm{sg}}^2.$$
4. *The plain objectives have bounded client drift, so that*
$$\sum_k a_k \|\nabla F_k(\omega) - \nabla F(\omega)\|_2^2 \le \sigma_{\mathrm{het}}^2.$$
5. *The semantic gradients are bounded, so that*
$$\|\nabla R_k^t(\omega)\|_2 \le G_R.$$

Lemma 29 (Regularized drift bound)**.** *Under Assumption 28,*

$$\sum_k a_k \|\nabla \widetilde{F}_k^t(\omega) - \nabla \widetilde{F}^t(\omega)\|_2^2 \le 2\sigma_{\mathrm{het}}^2 + 8\lambda^2 G_R^2. \tag{134}$$

Proof. Using
$$\nabla \widetilde{F}_k^t = \nabla F_k + \lambda \nabla R_k^t, \qquad \nabla \widetilde{F}^t = \nabla F + \lambda \nabla R^t,$$
we have
$$\begin{aligned} &\nabla \widetilde{F}_k^t(\omega) - \nabla \widetilde{F}^t(\omega) \\ &= (\nabla F_k(\omega) - \nabla F(\omega)) + \lambda \left(\nabla R_k^t(\omega) - \nabla R^t(\omega)\right). \end{aligned} \tag{135}$$
For any vectors u, v, $\|u+v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2$. Therefore
$$\begin{aligned} &\|\nabla \widetilde{F}_k^t - \nabla \widetilde{F}^t\|_2^2 \\ &\le 2\|\nabla F_k - \nabla F\|_2^2 + 2\lambda^2 \|\nabla R_k^t - \nabla R^t\|_2^2. \end{aligned} \tag{136}$$
By the bounded semantic-gradient assumption,
$$\begin{aligned} \|\nabla R^t(\omega)\|_2 &= \left\| \sum_j a_j \nabla R_j^t(\omega) \right\|_2 \\ &\le \sum_j a_j \|\nabla R_j^t(\omega)\|_2 \le G_R. \end{aligned}$$
Thus
$$\|\nabla R_k^t(\omega) - \nabla R^t(\omega)\|_2 \le 2G_R.$$
Multiplying Eq. (136) by a_k, summing over k, and using the plain drift assumption gives
$$\sum_k a_k \|\nabla \widetilde{F}_k^t - \nabla \widetilde{F}^t\|_2^2 \le 2\sigma_{\mathrm{het}}^2 + 2\lambda^2 (2G_R)^2,$$
which is Eq. (134).

Lemma 30 (Bias between regularized and plain gradients)**.** *For all ω and t,*

$$\|\nabla \widetilde{F}^t(\omega) - \nabla F(\omega)\|_2 \le \lambda G_R, \tag{137}$$

and

$$\|\nabla F(\omega)\|_2^2 \le 2\|\nabla \widetilde{F}^t(\omega)\|_2^2 + 2\lambda^2 G_R^2. \tag{138}$$

Proof. By definition,

$$\nabla \widetilde{F}^t(\omega) - \nabla F(\omega) = \lambda \nabla R^t(\omega).$$

As above,

$$\|\nabla R^t(\omega)\|_2 \le \sum_k a_k \|\nabla R_k^t(\omega)\|_2 \le G_R.$$

This proves Eq. (137). For the second claim, write

$$\nabla F = \nabla \widetilde{F}^t - (\nabla \widetilde{F}^t - \nabla F).$$

Then

$$\begin{aligned}\|\nabla F\|_2^2 &\le 2\|\nabla \widetilde{F}^t\|_2^2 + 2\|\nabla \widetilde{F}^t - \nabla F\|_2^2 \\ &\le 2\|\nabla \widetilde{F}^t\|_2^2 + 2\lambda^2 G_R^2.\end{aligned}$$

Lemma 31 (One-round local-SGD descent)**.** *Consider one communication round with H local stochastic-gradient steps, learning rate η_{lr}, and exact component-wise weighted averaging in the trainable factor/head coordinates. Under Assumption 28, if*

$$\eta_{\mathrm{lr}} \le \frac{1}{64 L_\lambda H},$$

then universal constants $A_1, A_2, A_3 > 0$ exist such that, with $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_0^t]$,

$$\begin{aligned}\mathbb{E}_t[F(\omega^{t+1})] \le F(\omega^t) &- \frac{\eta_{\mathrm{lr}} H}{4}\|\nabla F(\omega^t)\|_2^2 \\ &+ A_1 \eta_{\mathrm{lr}}^2 H L_\lambda \sigma_{\mathrm{sg}}^2 \\ &+ A_2 \eta_{\mathrm{lr}}^3 H^3 L_\lambda^2 \sigma_{\mathrm{het}}^2 \\ &+ A_3 \eta_{\mathrm{lr}} H \lambda^2 G_R^2.\end{aligned} \tag{139}$$

Proof. Condition on the state at the start of round t, and write

$$\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_0^t].$$

All expectations in this proof are conditional expectations $\mathbb{E}_t$ unless explicitly stated otherwise. Let

$$\omega_{k,0}^t = \omega^t, \qquad \omega_{k,h+1}^t = \omega_{k,h}^t - \eta_{\mathrm{lr}} g_{k,h}^t.$$

Exact weighted averaging gives

$$\begin{aligned}\omega^{t+1} &= \sum_k a_k \omega_{k,H}^t \\ &= \omega^t - \eta_{\mathrm{lr}} \sum_{h=0}^{H-1} \bar{g}_h^t, \qquad \bar{g}_h^t = \sum_k a_k g_{k,h}^t.\end{aligned} \tag{140}$$

Define

$$G_{k,h}^t = \nabla \widetilde{F}_k^t(\omega_{k,h}^t), \qquad G_t = \sum_k a_k G_{k,0}^t = \nabla \widetilde{F}^t(\omega^t),$$

$$G_t^F = \nabla F(\omega^t), \qquad \xi_{k,h}^t = g_{k,h}^t - G_{k,h}^t.$$

Let

$$\begin{aligned}D_t^{\mathrm{reg}} &= \sum_k a_k \|G_{k,0}^t - G_t\|_2^2, \\ V_t &= \sum_k a_k \|G_{k,0}^t\|_2^2 = \|G_t\|_2^2 + D_t^{\mathrm{reg}}.\end{aligned} \tag{141}$$

Set

$$U_h = \sum_k a_k \mathbb{E}\|\omega_{k,h}^t - \omega^t\|_2^2.$$

Since

$$\omega_{k,h}^t - \omega^t = -\eta_{\mathrm{lr}} \sum_{r=0}^{h-1} G_{k,r}^t - \eta_{\mathrm{lr}} \sum_{r=0}^{h-1} \xi_{k,r}^t,$$

the norm inequality, the Cauchy–Schwarz inequality, and the martingale property give

$$U_h \le 2\eta_{\mathrm{lr}}^2 h \sum_{r=0}^{h-1} \sum_k a_k \mathbb{E}\|G_{k,r}^t\|_2^2 + 2\eta_{\mathrm{lr}}^2 h \sigma_{\mathrm{sg}}^2. \tag{142}$$

By L_λ-smoothness,

$$\|G_{k,r}^t\|_2^2 \le 2\|G_{k,0}^t\|_2^2 + 2L_\lambda^2 \|\omega_{k,r}^t - \omega^t\|_2^2.$$

Hence

$$U_h \le 4\eta_{\mathrm{lr}}^2 h^2 V_t + 4\eta_{\mathrm{lr}}^2 h L_\lambda^2 \sum_{r=0}^{h-1} U_r + 2\eta_{\mathrm{lr}}^2 h \sigma_{\mathrm{sg}}^2. \tag{143}$$

Let $U_{\max} = \max_{0 \le h \le H} U_h$. Because $\eta_{\mathrm{lr}} L_\lambda H \le 1/64$, the recursive coefficient is at most $1/1024$. Absorbing it gives

$$U_{\max} \le 5\eta_{\mathrm{lr}}^2 H^2 V_t + 3\eta_{\mathrm{lr}}^2 H \sigma_{\mathrm{sg}}^2, \tag{144}$$

and therefore

$$\sum_{h=0}^{H-1} U_h \le 5\eta_{\mathrm{lr}}^2 H^3 V_t + 3\eta_{\mathrm{lr}}^2 H^2 \sigma_{\mathrm{sg}}^2. \tag{145}$$

Define

$$\delta_h^t = \mathbb{E}\left[\sum_k a_k (G_{k,h}^t - G_{k,0}^t)\right].$$

Jensen's inequality and smoothness imply $\|\delta_h^t\|_2^2 \le L_\lambda^2 U_h$. Thus

$$\sum_{h=0}^{H-1} \|\delta_h^t\|_2^2 \le 5\eta_{\mathrm{lr}}^2 H^3 L_\lambda^2 V_t + 3\eta_{\mathrm{lr}}^2 H^2 L_\lambda^2 \sigma_{\mathrm{sg}}^2. \tag{146}$$

Applying smoothness directly to the plain objective gives

$$\begin{aligned} F(\omega^{t+1}) \le F(\omega^t) &- \eta_{\mathrm{lr}} \sum_{h=0}^{H-1} \langle G_t^F, \bar{g}_h^t \rangle \\ &+ \frac{L\eta_{\mathrm{lr}}^2}{2} \left\| \sum_{h=0}^{H-1} \bar{g}_h^t \right\|_2^2. \end{aligned} \tag{147}$$

Conditional unbiasedness gives $\mathbb{E}[\bar{g}_h^t] = G_t + \delta_h^t$. Since

$$G_t = G_t^F + \lambda \nabla R^t(\omega^t), \qquad \|\nabla R^t(\omega^t)\|_2 \le G_R,$$

Young's inequality yields

$$-\langle G_t^F, G_t \rangle \le -\frac{3}{4}\|G_t^F\|_2^2 + \lambda^2 G_R^2. \tag{148}$$

Also,

$$-\langle G_t^F, \delta_h^t \rangle \le \frac{1}{8}\|G_t^F\|_2^2 + 2\|\delta_h^t\|_2^2.$$

Therefore

$$\begin{aligned} &- \eta_{\mathrm{lr}} \sum_{h=0}^{H-1} \mathbb{E}\langle G_t^F, \bar{g}_h^t \rangle \\ &\le -\frac{5\eta_{\mathrm{lr}} H}{8}\|G_t^F\|_2^2 + \eta_{\mathrm{lr}} H \lambda^2 G_R^2 + 2\eta_{\mathrm{lr}} \sum_{h=0}^{H-1} \|\delta_h^t\|_2^2. \end{aligned} \tag{149}$$

Put

$$\bar{G}_h^t = \sum_k a_k G_{k,h}^t, \qquad \bar{\xi}_h^t = \sum_k a_k \xi_{k,h}^t.$$

The conditional independence and martingale assumptions imply

$$\mathbb{E}\left\| \sum_{h=0}^{H-1} \bar{\xi}_h^t \right\|_2^2 \le H\sigma_{\mathrm{sg}}^2.$$

For the drift term,

$$\begin{aligned} \mathbb{E}\left\| \sum_{h=0}^{H-1} \bar{G}_h^t \right\|_2^2 &\le 2H^2\|G_t\|_2^2 + 2H \sum_{h=0}^{H-1} \mathbb{E}\|\bar{G}_h^t - G_t\|_2^2 \\ &\le 2H^2\|G_t\|_2^2 + 2HL_\lambda^2 \sum_{h=0}^{H-1} U_h. \end{aligned} \tag{150}$$

Combining with Eq. (145) gives

$$\begin{aligned} \mathbb{E}\left\| \sum_{h=0}^{H-1} \bar{g}_h^t \right\|_2^2 &\le 4H^2\|G_t\|_2^2 + 2H\sigma_{\mathrm{sg}}^2 \\ &\quad + 20\eta_{\mathrm{lr}}^2 H^4 L_\lambda^2 V_t \\ &\quad + 12\eta_{\mathrm{lr}}^2 H^3 L_\lambda^2 \sigma_{\mathrm{sg}}^2. \end{aligned} \tag{151}$$

The factor η_{lr}^2 in the client-drift contribution follows from the local displacement in Eq. (143).

Lemma 29 gives

$$D_t^{\mathrm{reg}} \le 2\sigma_{\mathrm{het}}^2 + 8\lambda^2 G_R^2,$$

and

$$\|G_t\|_2^2 \le 2\|G_t^F\|_2^2 + 2\lambda^2 G_R^2.$$

Consequently,

$$\begin{aligned} V_t &= \|G_t\|_2^2 + D_t^{\mathrm{reg}} \\ &\le 2\|G_t^F\|_2^2 + 2\sigma_{\mathrm{het}}^2 + 10\lambda^2 G_R^2. \end{aligned} \tag{152}$$

Substitute Eqs. (146), (149), and (151) into Eq. (147), and then use Eq. (152). Put $x = \eta_{\mathrm{lr}} L_\lambda H$. Using $L \le L_\lambda$, the positive gradient-dependent contribution from the first-order drift is at most

$$20x^2\eta_{\mathrm{lr}} H\|G_t^F\|_2^2,$$

the leading quadratic-smoothness contribution is at most

$$4x\eta_{\mathrm{lr}} H\|G_t^F\|_2^2,$$

and its local-displacement contribution is at most

$$20x^3\eta_{\mathrm{lr}} H\|G_t^F\|_2^2.$$

For $x \le 1/64$,

$$4x + 20x^2 + 20x^3 < \frac{3}{8}.$$

Their sum is therefore absorbed by the negative term $-5\eta_{\mathrm{lr}} H\|G_t^F\|_2^2/8$ in Eq. (149), leaving at least $-\eta_{\mathrm{lr}} H\|G_t^F\|_2^2/4$. The remaining terms have orders

$$\eta_{\mathrm{lr}}^2 H L_\lambda \sigma_{\mathrm{sg}}^2, \qquad \eta_{\mathrm{lr}}^3 H^3 L_\lambda^2 \sigma_{\mathrm{het}}^2, \qquad \eta_{\mathrm{lr}} H\lambda^2 G_R^2.$$

Collecting numerical coefficients into A_1, A_2, A_3 proves Eq. (139).

Theorem 32 (Reference local-SGD convergence bound)**.** *Suppose Assumption 28 holds. The server uses exact weighted averaging. Each round has H local steps and learning rate η_{lr}. If $\eta_{\mathrm{lr}} \le 1/(64L_\lambda H)$, then*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\omega^t)\|_2^2 \\ \le \frac{4(F(\omega^0)-F^\star)}{\eta_{\mathrm{lr}}HT} + C_1\eta_{\mathrm{lr}}L_\lambda\sigma_{\mathrm{sg}}^2 \\ + C_2\eta_{\mathrm{lr}}^2L_\lambda^2H^2\sigma_{\mathrm{het}}^2 \\ + C_3\lambda^2G_R^2,\end{aligned} \tag{153}$$

for universal constants $C_1, C_2, C_3 > 0$. For any fixed $c > 0$, choose

$$\eta_{\mathrm{lr},T} = \min\left\{\frac{c}{\sqrt{HT}}, \frac{1}{64L_\lambda H}\right\}. \tag{154}$$

This choice always satisfies the required stability condition. When

$$T \ge 4096c^2L_\lambda^2H, \tag{155}$$

the first branch is active and

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\omega^t)\|_2^2 \\ \le C_0'\frac{F(\omega^0)-F^\star}{c\sqrt{HT}} \\ + C_1'\frac{cL_\lambda\sigma_{\mathrm{sg}}^2}{\sqrt{HT}} \\ + C_2'\frac{c^2L_\lambda^2H\sigma_{\mathrm{het}}^2}{T} \\ + C_3'\lambda^2G_R^2,\end{aligned} \tag{156}$$

for constants $C_0', C_1', C_2', C_3' > 0$ independent of H and T.

Proof. Apply the tower property to the conditional inequality in Lemma 31, and then sum it over $t = 0, \dots, T-1$:

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{\eta_{\mathrm{lr}}H}{4}\mathbb{E}\|\nabla F(\omega^t)\|_2^2 \\ \le F(\omega^0) - \mathbb{E}[F(\omega^T)] \\ + A_1T\eta_{\mathrm{lr}}^2HL_\lambda\sigma_{\mathrm{sg}}^2 \\ + A_2T\eta_{\mathrm{lr}}^3H^3L_\lambda^2\sigma_{\mathrm{het}}^2 \\ + A_3T\eta_{\mathrm{lr}}H\lambda^2G_R^2.\end{aligned} \tag{157}$$

Since $F(\omega^T) \ge F^\star$,

$$F(\omega^0) - \mathbb{E}[F(\omega^T)] \le F(\omega^0) - F^\star.$$

Divide Eq. (157) by $\eta_{\mathrm{lr}}HT/4$. This gives Eq. (153) after absorbing numerical constants into C_1, C_2, C_3. The choice in Eq. (154) is no larger than $1/(64L_\lambda H)$, so Eq. (153) applies for all T. If Eq. (155) holds, then

$$\frac{c}{\sqrt{HT}} \le \frac{1}{64L_\lambda H},$$

so $\eta_{\mathrm{lr},T} = c/\sqrt{HT}$. Consequently,

$$\frac{1}{\eta_{\mathrm{lr},T}HT} = O\left(\frac{1}{\sqrt{HT}}\right), \qquad \eta_{\mathrm{lr},T} = O\left(\frac{1}{\sqrt{HT}}\right),$$

and

$$\eta_{\mathrm{lr},T}^2 H^2 = O\left(\frac{H}{T}\right).$$

Substituting these relations into Eq. (153) proves Eq. (156). For smaller T, the exact non-asymptotic statement remains Eq. (153) with the second branch of Eq. (154).

C.8 Transfer to the implemented local optimizer

The preceding theorem is a fixed-H, plain-local-SGD reference result. Fix any reference step count H satisfying its step-size condition. The experiment instead uses momentum, two parameter-group rates, and a client-dependent number $H_{k,t}$ of mini-batch steps; discrepancies in these step counts are included in the endpoint residual below.

Let $g_{k,h}^{\mathrm{act},t}$ be the implemented update direction after the common gradient clipping and weight-decay operations. With momentum $m = 0.9$, base learning rate $\eta_{\mathrm{lr}} = 0.01$, and

$$P = \mathrm{diag}(1 \text{ on LoRA/head}, \ 0.1 \text{ on LayerNorm}),$$

the local recursion is

$$\begin{aligned} v_{k,h+1}^t &= m v_{k,h}^t + g_{k,h}^{\mathrm{act},t}, \\ \omega_{k,h+1}^{\mathrm{act},t} &= \omega_{k,h}^{\mathrm{act},t} - \eta_{\mathrm{lr}} P v_{k,h+1}^t. \end{aligned} \tag{158}$$

For $q \geq 1$, put $c_q(m) = (1 - m^q)/(1 - m)$. Directly unrolling Eq. (158) gives the exact endpoint identity

$$\begin{aligned} \omega_{k,H_{k,t}}^{\mathrm{act},t} = \omega^t - \eta_{\mathrm{lr}} P \Bigg[& \frac{m(1 - m^{H_{k,t}})}{1 - m} v_{k,0}^t \\ & + \sum_{h=0}^{H_{k,t}-1} c_{H_{k,t}-h}(m) g_{k,h}^{\mathrm{act},t} \Bigg]. \end{aligned} \tag{159}$$

The first term vanishes when the local optimizer buffer is initialized to zero. In contrast, let $\omega_{k,H}^{\mathrm{ref},t}$ be the H-step plain-SGD endpoint in Lemma 31, started from the same ω^t and driven by its coupled reference gradients. Define the exact weighted endpoint discrepancy

$$\mathcal{E}_{\mathrm{opt},t} = \left\| \sum_k a_k \omega_{k,H_{k,t}}^{\mathrm{act},t} - \sum_k a_k \omega_{k,H}^{\mathrm{ref},t} \right\|_2. \tag{160}$$

Equation (159) makes $\mathcal{E}_{\mathrm{opt},t}$ a pathwise quantity determined by the actual momentum, group rates, mini-batch sequence, and client-specific step counts.

Let $\mathcal{L}_t$ be the closed line segment between the two weighted endpoints in Eq. (160), and define the pathwise finite quantity

$$G_{\mathrm{opt},t} = \max_{\omega \in \mathcal{L}_t} \|\nabla F(\omega)\|_2. \tag{161}$$

The maximum exists because L-smoothness makes the gradient continuous and $\mathcal{L}_t$ is compact.

Lemma 33 (Objective transfer for the implemented local optimizer)**.** *Under the L-smoothness condition in Assumption 28,*

$$\begin{aligned} F(\omega_{\mathrm{fac,act}}^{t+1}) &\leq F(\omega_{\mathrm{fac,ref}}^{t+1}) + \delta_t^{\mathrm{opt}}, \\ \delta_t^{\mathrm{opt}} &= G_{\mathrm{opt},t} \mathcal{E}_{\mathrm{opt},t} + \frac{L}{2} \mathcal{E}_{\mathrm{opt},t}^2. \end{aligned} \tag{162}$$

Proof. Apply the L-smoothness inequality between the two endpoints. the Cauchy–Schwarz inequality bounds the linear term by $G_{\mathrm{opt},t} \mathcal{E}_{\mathrm{opt},t}$, and the quadratic term is $L\mathcal{E}_{\mathrm{opt},t}^2/2$.

C.9 IMPLEMENTED SERVER MAP IN FUNCTIONAL COORDINATES

The comparison below starts from the same factor-averaged actual client endpoints and isolates only the change introduced by the implemented server map. Theorem 32 uses a reference server that averages the trainable LoRA factors and head/norm coordinates component-wise. The implemented system first reaches the actual momentum/group-rate client endpoints, averages the represented dense LoRA increments, reconstructs them through the averaged first factor by Eq. (7), rebalances the resulting factors, and anchors the classifier. We compare the factor-averaged actual endpoint with this server output in one functional model coordinate.

For every LoRA-injected projection ℓ, let

$$\Delta_{k,\ell}^{t+1} = (A_{k,\ell}^{t+1}, B_{k,\ell}^{t+1}).$$

The factor-averaged reference state and represented increment are

$$\begin{aligned} \Delta_{\mathrm{fac},\ell}^{t+1} &= \sum_k a_k^t \Delta_{k,\ell}^{t+1}, \\ X_{\mathrm{fac},\ell}^{t+1} &= \mathrm{Dense}(\Delta_{\mathrm{fac},\ell}^{t+1}). \end{aligned} \tag{163}$$

The implemented function-space target and ridge-constrained reconstruction are

$$\begin{aligned} X_{\mathrm{fun},\ell}^{t+1} &= \sum_k a_k^t \,\mathrm{Dense}(\Delta_{k,\ell}^{t+1}), \\ \bar{A}_\ell^{t+1} &= \sum_k a_k^t A_{k,\ell}^{t+1}, \\ \bar{B}_\ell^{t+1} &= \arg\min_B \left\{ \|B\bar{A}_\ell^{t+1} - X_{\mathrm{fun},\ell}^{t+1}\|_F^2 + \lambda_{\mathrm{r}} \|B\|_F^2 \right\}, \\ X_{\mathrm{impl},\ell}^{t+1} &= \bar{B}_\ell^{t+1} \bar{A}_\ell^{t+1}. \end{aligned} \tag{164}$$

The rebalancing in Eq. (20) preserves $X_{\mathrm{impl},\ell}^{t+1}$, since this product already has rank at most r. Define

$$\begin{aligned} \mathcal{H}_{\mathrm{coord},t}^2 &= \sum_\ell \|X_{\mathrm{fac},\ell}^{t+1} - X_{\mathrm{fun},\ell}^{t+1}\|_F^2, \\ \mathcal{H}_{\mathrm{rec},t}^2 &= \sum_\ell \|X_{\mathrm{fun},\ell}^{t+1} - X_{\mathrm{impl},\ell}^{t+1}\|_F^2, \\ \mathcal{H}_{\mathrm{map},t}^2 &= \sum_\ell \|X_{\mathrm{fac},\ell}^{t+1} - X_{\mathrm{impl},\ell}^{t+1}\|_F^2. \end{aligned} \tag{165}$$

The triangle inequality gives

$$\mathcal{H}_{\mathrm{map},t} \le \mathcal{H}_{\mathrm{coord},t} + \mathcal{H}_{\mathrm{rec},t}. \tag{166}$$

Let $\widetilde{W}^{t+1}$ be the exactly averaged classifier before anchoring and $\bar{W}^{t+1}$ the implemented classifier. Define

$$\mathcal{H}_{\mathrm{anc},t}^2 = \sum_{c=1}^{C} \|\bar{W}_c^{t+1} - \widetilde{W}_c^{t+1}\|_2^2, \tag{167}$$

and

$$\mathcal{H}_{\mathrm{srv},t} = \left(\mathcal{H}_{\mathrm{map},t}^2 + \mathcal{H}_{\mathrm{anc},t}^2\right)^{1/2}. \tag{168}$$

Assumption 34 (Functional-coordinate server smoothness)**.** *View the model objective as a function* $\mathfrak{F}(\Theta)$ *of*

$$\Theta = (\{X_\ell\}_\ell, W, b, \psi),$$

where X_ℓ *is the represented dense LoRA increment of projection* ℓ*. On each segment joining the factor-averaged reference model to the implemented server model,* $\mathcal{F}$ *is* L_{fun}*-smooth and*

$$\|\nabla_\Theta \mathfrak{F}(\Theta)\|_{\mathrm{fun}} \le G_{\mathrm{fun}},$$

under

$$\|\Theta\|_{\mathrm{fun}}^2 = \sum_\ell \|X_\ell\|_F^2 + \|W\|_F^2 + \|b\|_2^2 + \|\psi\|_2^2.$$

Lemma 35 (Objective change caused by the implemented server map)**.** *Under Assumption 34,*

$$\begin{aligned} F(\omega_{\mathrm{impl}}^{t+1}) &\le F(\omega_{\mathrm{fac,act}}^{t+1}) + \delta_t^{\mathrm{srv}}, \\ \delta_t^{\mathrm{srv}} &= G_{\mathrm{fun}}\mathcal{H}_{\mathrm{srv},t} + \frac{L_{\mathrm{fun}}}{2}\mathcal{H}_{\mathrm{srv},t}^2. \end{aligned} \tag{169}$$

Proof. Let $\Theta_{\mathrm{fac,act}}^{t+1}$ and $\Theta_{\mathrm{impl}}^{t+1}$ be the represented functional coordinates of the two endpoints. Their product-norm difference is $\mathcal{H}_{\mathrm{srv},t}$. Functional smoothness gives

$$\begin{aligned} \mathfrak{F}(\Theta_{\mathrm{impl}}^{t+1}) \le\ & \mathfrak{F}(\Theta_{\mathrm{fac,act}}^{t+1}) \\ & + \left\langle \nabla_\Theta \mathfrak{F}(\Theta_{\mathrm{fac,act}}^{t+1}), \Theta_{\mathrm{impl}}^{t+1} - \Theta_{\mathrm{fac,act}}^{t+1} \right\rangle_{\mathrm{fun}} \\ & + \frac{L_{\mathrm{fun}}}{2}\mathcal{H}_{\mathrm{srv},t}^2. \end{aligned} \tag{170}$$

The Cauchy–Schwarz inequality bounds the inner product by $G_{\mathrm{fun}}\mathcal{H}_{\mathrm{srv},t}$. Since $\mathfrak{F}(\Theta)$ and $F(\omega)$ are the same model objective under two coordinate descriptions, Eq. (169) follows.

Theorem 36 (Residual-aware stationarity bound for the implemented optimizer and server)**.** *Suppose the assumptions of Theorem 32 and Assumption 34 hold, and $\eta_{\mathrm{lr}} \le 1/(64L_\lambda H)$. Then the actual iterates satisfy*

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\omega^t)\|_2^2 \\ &\quad \le \frac{4(F(\omega^0) - F^\star)}{\eta_{\mathrm{lr}} HT} + C_1 \eta_{\mathrm{lr}} L_\lambda \sigma_{\mathrm{sg}}^2 \\ &\qquad + C_2 \eta_{\mathrm{lr}}^2 L_\lambda^2 H^2 \sigma_{\mathrm{het}}^2 + C_3 \lambda^2 G_R^2 \\ &\qquad + \frac{4}{\eta_{\mathrm{lr}} HT}\sum_{t=0}^{T-1} \mathbb{E}[\delta_t^{\mathrm{opt}} + \delta_t^{\mathrm{srv}}]. \end{aligned} \tag{171}$$

Proof. At round t, apply Lemma 31 to the coupled factor-averaged plain-SGD reference endpoint. Then apply Lemma 33 to reach the factor-averaged actual local-optimizer endpoint, followed by Lemma 35 to reach the implemented server output. This adds $\delta_t^{\mathrm{opt}} + \delta_t^{\mathrm{srv}}$ to the one-round descent bound. The implemented output is the start of the next round, so summing over $t = 0, \dots, T-1$ telescopes the actual objective values into

$$\begin{aligned} &\sum_{t=0}^{T-1} \frac{\eta_{\mathrm{lr}} H}{4} \mathbb{E}\|\nabla F(\omega^t)\|_2^2 \\ &\quad \le F(\omega^0) - \mathbb{E}[F(\omega^T)] \\ &\qquad + A_1 T \eta_{\mathrm{lr}}^2 H L_\lambda \sigma_{\mathrm{sg}}^2 + A_2 T \eta_{\mathrm{lr}}^3 H^3 L_\lambda^2 \sigma_{\mathrm{het}}^2 \\ &\qquad + A_3 T \eta_{\mathrm{lr}} H \lambda^2 G_R^2 + \sum_{t=0}^{T-1} \mathbb{E}[\delta_t^{\mathrm{opt}} + \delta_t^{\mathrm{srv}}]. \end{aligned} \tag{172}$$

Use $F(\omega^T) \ge F^\star$ and divide by $\eta_{\mathrm{lr}} HT/4$ to obtain Eq. (171).

Implementation residual. Theorem 36 establishes a finite-horizon stationarity bound that accounts for deviations between the plain-SGD reference and the implemented optimizer and server map. Fixed classifier anchoring and optimizer mismatch may introduce nonvanishing residuals. Consequently, the bound characterizes stationarity up to the implementation and semantic-regularization residuals rather than guaranteeing exact asymptotic stationarity.

C.10 SCORE-WEIGHTED READOUT

At inference, client k combines the linear and prototype probabilities through a single weight,

$$q_k^{(\alpha)}(x) = \alpha q_{\mathrm{lin},k}(x) + (1-\alpha) q_{\mathrm{proto},k}(x), \tag{173}$$

starting from the score-derived prior $\widehat{\alpha}_k = \mathrm{clamp}(1 - \beta \bar{s}_k, \alpha_{\min}, 1)$ of Eq. (16), where $\bar{s}_k = \sum_c (n_{k,c}/n_k) s_c^\star$ is the support-weighted mean of the refreshed scores over the classes client k observes. The candidate set is

$$\mathcal{A}_k = \Big\{ \mathrm{clamp}(\widehat{\alpha}_k - 0.1, \alpha_{\min}, 1),\ \widehat{\alpha}_k,\ \mathrm{clamp}(\widehat{\alpha}_k + 0.1, \alpha_{\min}, 1),\ 1 \Big\}, \tag{174}$$

The candidate set includes a local correction around the score-derived prior and the linear endpoint. Each candidate is evaluated on the client's training partition, with ties resolved in favor of the largest α,

$$\alpha_k^\star = \max \left\{ \alpha \in \mathcal{A}_k : \mathrm{Acc}_{\mathcal{D}_k}(q_k^{(\alpha)}) = \max_{a \in \mathcal{A}_k} \mathrm{Acc}_{\mathcal{D}_k}(q_k^{(a)}) \right\}. \tag{175}$$

If the final bank is empty, the implemented fallback is $\alpha_k^\star = 1$ and Eq. (175) is not evaluated.

Proposition 37 (Readout-prior monotonicity and candidate-set optimality)**.** *The prior $\widehat{\alpha}_k$ is nonincreasing in $\bar{s}_k$, equals 1 when $\bar{s}_k = 0$, and satisfies $\widehat{\alpha}_k \geq \alpha_{\min}$ for every $\bar{s}_k \in [0,1)$. Since $1 \in \mathcal{A}_k$,*

$$\mathrm{Acc}_{\mathcal{D}_k}(q_k^{(\alpha_k^\star)}) \geq \mathrm{Acc}_{\mathcal{D}_k}(q_k^{(1)}). \tag{176}$$

Proof. The map $s \mapsto 1 - \beta s$ is affine with slope $-\beta \leq 0$ and clamping to $[\alpha_{\min}, 1]$ is nondecreasing, so the composition is nonincreasing in $\bar{s}_k$. At $\bar{s}_k = 0$ the unclamped value is $1 \in [\alpha_{\min}, 1]$, so the clamp returns 1. The lower bound is immediate from the clamp, and $s_c^\star \in [0,1)$ by Eq. (12) with weights summing to one gives $\bar{s}_k \in [0,1)$. For the second claim, $1 \in \mathcal{A}_k$ makes the pure linear readout feasible in the same maximization that defines $\alpha_k^\star$, and a maximum over a set is at least the value at any element of that set.

Remark 38. *The optimality in Proposition 37 is defined on $\mathcal{D}_k$. Its test-set contribution is evaluated through the* linear-only *ablation in Table 2.*

C.11 SUMMARY OF THE PROOF CHAIN

The proof chain has four stages. First, Lemma 4, Remark 5, and Lemma 6 separate the fixed-map split analysis from the fitted-encoder reuse correction. Lemma 7 and Lemmas 9–12 then establish the prototype-noise scale, while Lemma 13 and Theorem 14 give the expectation-level calibrated dispersion result; Lemma 15 isolates the additional temporal lag from smoothing.

Second, Theorem 17, Lemma 18, and Theorem 19 relate $D_c^\star$ to one- and multi-prototype population approximation. Corollary 20 gives a sufficient condition for a strict Euclidean multi-prototype gain, and Proposition 21 bounds the change from the reliability-weighted population objective to the support-weighted objective used by the bank. The empirical-bank discussion then separates this population reference from the finite-iteration spherical clustering used in the implementation.

Third, Lemma 22 and Proposition 23 connect the stabilized positive-part/EMA score to its population target, while Propositions 24–27 establish the stated monotonicity or sufficient-direction properties of capacity, classifier anchoring, semantic gating, and the inference prior.

Finally, Lemmas 29–31 and Theorem 32 give the reference local-SGD stationarity bound. The exact momentum endpoint identity Eq. (159), Lemma 33, Lemma 35, and Theorem 36 transfer that reference to the implemented optimizer, functional aggregation, ridge-constrained reconstruction, and classifier anchoring with explicit implementation residuals. Proposition 37 establishes readout-prior monotonicity and candidate-set optimality under the local selection criterion.