TACO: TEXT-ANCHORED CLASSIFICATION WITH STRUCTURAL COVARIANCE TRANSFER FOR ONE-SHOT FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional federated learning (FL) requires multiple communication rounds and incurs high overhead, while One-shot FL (OFL) mitigates this burden by restricting interaction to a single round. However, most existing OFL methods either require clients to upload prototypes or mandate homogeneous client architectures. The former is prone to privacy leakage risks, while the latter greatly limits the scalability. The statistics-based OFL method constructs a classifier using per-class sufficient statistics derived from features extracted by a frozen pre-trained encoder, achieving partition invariance, that is, identical accuracy regardless of data distribution across clients. Nevertheless, current statistics-based OFL methods estimate each class independently, so that when per-class samples are scarce, covariances are regularized by shrinking toward an isotropic target that discards the covariance structure shared across classes. Since all classes pass through the same frozen pre-trained encoder, their covariances share common variation patterns, a source of structural information that isotropic shrinkage entirely wastes, leading to degraded performance as the number of classes grows. To address these limitations, we propose **TACO** (**T**ext-**A**nchored **C**lassification with Structural C**o**variance Transfer), a training-free OFL framework built on three components. First, *Structural Covariance Transfer* regularizes each class covariance toward the average of all other classes rather than a scaled identity, directly exploiting the shared variation patterns of the frozen pre-trained encoder, and its gains grow as per-class estimation becomes harder (up to $+5.57\%$ on Tiny-ImageNet). Second, *dual-space QDA/LDA fusion* uses text anchors generated by CLIP to construct a complementary text-space classifier and fuses it with the visual-space classifier, integrating discriminative information unavailable from either space alone. Third, *confidence-weighted test-time augmentation* addresses the bottleneck that training-side parameter estimation has limited room for improvement while per-sample feature noise at test time remains, providing orthogonal gains via multi-view confidence-weighted aggregation. Extensive experiments demonstrate that TACO outperforms or matches all SOTA OFL methods. The code of this work will be released upon acceptance.

1 INTRODUCTION

Federated learning (FL) enables multiple clients to collaboratively train models without sharing raw data, and has been widely adopted in privacy-sensitive applications (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020). However, conventional FL requires multiple communication rounds, incurring substantial overhead and repeated privacy exposure. For instance, on CIFAR-10 with non-IID settings, FedAvg (McMahan et al., 2017) achieves below 35% accuracy in a single round, while exceeding 85% demands over 50 rounds (Turazza et al., 2026). Each additional round imposes further bandwidth costs and widens the window for potential privacy leakage.

One-shot federated learning (OFL) limits communication to a single round (Liu et al., 2025), but demands higher quality from each client's transmission. Existing OFL methods typically require clients to upload prototypes or trained model parameters. However, the former poses a substantial risk of privacy leakage, while the latter requires identical model architectures across all clients, severely limiting scalability. Among OFL paradigms, the *statistics-based* method stands out for its structural

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.

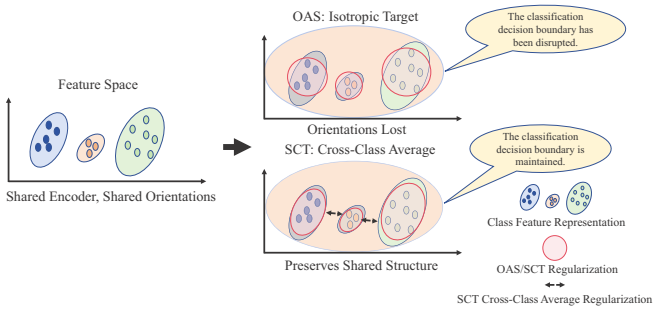

Figure 1: OAS vs. SCT. OAS shrinks toward a scaled identity, discarding directional structure. SCT instead targets the cross-class average, preserving shared structure.

advantage. Each client computes per-class sufficient statistics from features extracted by a frozen pre-trained encoder and uploads them to the server, and the server constructs classifiers without any gradient-based training (Legate et al., 2023; Fanì et al., 2024; Guan et al., 2025; Turazza et al., 2026; He et al., 2025). Since the summation operation is order-agnostic, the resulting model at the server is *partition-invariant*, meaning that mathematically the outcome remains identical regardless of how data are distributed across clients. This gives these methods inherent robustness against data heterogeneity, a guarantee that is typically absent in *train-then-merge* methods (Zeng et al., 2025).

Nevertheless, current *statistics-based* methods estimate each class in isolation. Per-class means and covariances are computed independently, and each class covariance is regularized without reference to any other class. When the number of per-class samples is small relative to the feature dimension, the sample covariance becomes rank-deficient and unreliable. Existing methods address this via Oracle Approximating Shrinkage (OAS) (Chen et al., 2010), which interpolates each class covariance toward a scaled identity. However, because all classes share a common frozen pre-trained encoder, their covariances inherit common variation patterns from the encoder's learned representations. OAS discards this shared directional structure entirely, introducing unnecessary bias. The covariances of other classes naturally capture this structure and could serve as a far more informative regularization target, yet no existing OFL method leverages such cross-class covariance information.

Additionally, recent works widely adopt frozen CLIP (Radford et al., 2021) for *statistics-based* OFL (He et al., 2025; Zhang et al., 2026; Goswami et al., 2025), yet none of them tap into the potential of CLIP's text encoder. The text branch of CLIP produces class-name embeddings for each category that carry rich semantic information, and these *text anchors* naturally encode inter-class semantic relationships and define a classification space complementary to the visual space. Ignoring this text modality confines the classifier to a single visual feature space, forgoing the additional discriminative information that the text side can offer. Meanwhile, on the training side, because sufficient statistics already aggregate over all available samples, class-level representations approach saturation, leaving extremely limited room for further optimization. The performance bottleneck has therefore shifted from training to test time. Since each query sample is encoded by a single forward pass, its representation inevitably absorbs input-level variability such as crops, illumination shifts, and compression artifacts. This per-sample noise propagates directly into classification decisions and constitutes the dominant source of residual error.

To address these limitations, this paper proposes **TACO** (**T**ext-**A**nchored **C**lassification with Structural **Co**variance Transfer), a training-free OFL framework. Our main contributions are summarized as follows:

- **Structural Covariance Transfer (SCT).** Because all classes share a common frozen pre-trained encoder, their covariances exhibit shared variation patterns that conventional isotropic shrinkage fails to preserve. As illustrated in Figure 1, OAS shrinks each class covariance toward a scaled identity, discarding all directional structure. In contrast, SCT regularizes each estimate toward the cross-class average covariance with a sample-size-adaptive mixing weight, so that data-scarce classes borrow more shared structure while data-rich classes retain their own estimates.

- **Dual-Space QDA/LDA Fusion.** Text anchors generated by CLIP project visual features into a C-dimensional text space, filtering out class-irrelevant dimensions from the original D-dimensional visual space. We build independent QDA classifiers, each modeling a class with its own Gaussian distribution in both spaces and fuse their log-posteriors. This integrates two structurally complementary views of the data. The visual space captures fine-grained per-class variation, while the text space emphasizes inter-class semantic separation.
- **Confidence-Weighted Test-Time Augmentation (CTTA).** To mitigate per-sample feature noise, CTTA generates multiple augmented views for each test sample and aggregates their predictions through confidence-based weighting, effectively denoising the noisy single-pass estimate. The resulting gains are orthogonal to all training-side components and grow as estimation difficulty increases.
- **New State-Of-The-Art (SOTA) without Any Training.** TACO achieves 92.06%, 76.69%, and 71.00% on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, establishing new SOTA results among OFL methods. It outperforms or matches both *train-then-merge* approaches and prior *statistics-based* methods, while requiring no gradient-based optimization and remaining partition-invariant. On Tiny-ImageNet, classifier construction completes in under 10 seconds on a single GPU, and a text-only variant reduces the CIFAR-100 upload to 1.97 MiB per client ($\sim 25\times$ smaller than full-space statistics) while remaining competitive (see Appendix G).

2 Related Works

OFL restricts client-server interaction to a single communication round, trading multi-round refinement for communication efficiency (Liu et al., 2025). The core challenge is to maximize the utility of each client's single upload. Depending on what clients transmit, existing methods fall into three paradigms.

Train-then-Merge. In this paradigm, clients independently train local models on their private data and upload parameters, features, logits or prototypes for server-side aggregation. Representative approaches include parameter averaging (McMahan et al., 2017; Su et al., 2023), Fisher-weighted merging (Jhunjhunwala et al., 2024), data-free distillation (Zhang et al., 2022; Lin et al., 2020), ensemble co-boosting (Dai et al., 2024), causal fusion (Tang et al., 2024), Bayesian posterior aggregation (Liu et al., 2024; Talpini et al., 2025), ensembling (Allouah et al., 2024; Zeng et al., 2024) and feature alignment (Zeng et al., 2025). These methods face a fundamental bottleneck under non-IID data, namely *representation inconsistency*, where feature representations and decision boundaries diverge across clients. Zeng et al. (2025) identify this inconsistency as the root cause of performance degradation, with accuracy dropping sharply as data heterogeneity increases.

Foundation Model with Generation. Clients avoid local training by extracting sample features using a frozen base model, fit lightweight generative models to local feature distributions, and send the generative parameters to the server, which then synthesizes features or data for classifier training. FedPFT (Beitollahi et al., 2025) fits Gaussian mixture models to per-class features and sends parameters to the server, which samples synthetic features and trains a classification head. FedBiP (Chen et al., 2025) uses client-side frozen features to drive latent diffusion models and generates synthetic images for server-side distillation. FedD3 (Song et al., 2023) distills local datasets into compact synthetic representations and uploads them to the server for one-shot centralized training. FALCON (Liu et al., 2026) fits an autoregressive generator to the token sequences of each client and uploads synthesized sequences for server-side distillation. While these methods avoid model inconsistency by sharing a common encoder, they incur extra computational overhead from generative modeling and sampling on the server. Moreover, the quality of the synthetic data depends critically on the fidelity of the generative model, which may be unreliable when per-class samples are scarce. Empirically, generation-based methods remain substantially below statistics-based approaches (Turazza et al., 2026).

Foundation Model with Statistics. The third paradigm uploads compact per-class sufficient statistics from the sample features obtained by the frozen pre-trained encoder and analytically constructs classifiers, avoiding local training and generative modeling. This line of work has evolved rapidly, from nearest-centroid classification (Legate et al., 2023), to ridge regression with second moments (Fed3R (Fanì et al., 2024)), to shared-covariance Gaussian classifiers (Guan et al., 2025), and finally

to the full Gaussian discriminant family (NB/LDA/QDA) with Fisher-subspace trainable heads (GH-OFL (Turazza et al., 2026)). AFL (He et al., 2025) derives analytic solutions for pre-trained encoders. DeepAFL (Tang et al., 2026) extends analytic federated learning to deep residual networks through layer-wise closed-form optimization. SAFLe (Bacellar et al., 2026) adds a bucketed embedding head to AFL within a single round. FedCOF (Goswami et al., 2025) estimates covariances solely from mean distributions. Although recent works adopt frozen CLIP features (He et al., 2025; Zhang et al., 2026; Goswami et al., 2025), they do not exploit text embeddings to construct complementary classification spaces. Regarding covariance modeling, classical shrinkage (Chen et al., 2010) and Ledoit-Wolf estimators (Ledoit & Wolf, 2004) interpolate between the sample covariance and a scaled identity, an isotropic target that ignores inter-class relationships. Few-shot methods borrow statistics from base classes for covariance calibration (Yang et al., 2021), but this idea remains unused in OFL. Most statistics-based OFL methods avoid per-class covariance estimation altogether, and the few that do regularize their estimates toward generic isotropic targets, as in OAS (Chen et al., 2010), while ignoring inter-class covariance structure. SCT is, to our knowledge, the first to use cross-class covariance as the shrinkage target in OFL, blending each class's covariance with the average of others to exploit the shared variation patterns of the frozen pre-trained encoder.

3 METHOD

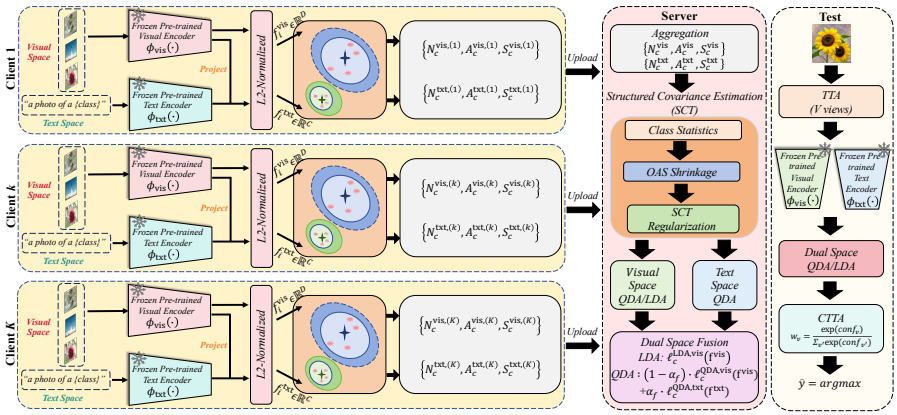

Figure 2: Overview of TACO. **Client side:** each client uploads per-class sufficient statistics for both spaces. **Server side:** statistics from K clients are aggregated with OAS shrinkage and SCT. For $C > 20$, dual-space QDA/LDA classifiers are built and their log-posteriors are fused with a mixing weight α_f. For $C \leq 20$, only a visual LDA classifier is used. **Test:** TTA generates V augmented views, and CTTA aggregates predictions with confidence weights w_v to produce $\hat{y}$.

Problem Formulation. TACO operates in a standard OFL setting with K clients and a central server. All clients share a frozen pre-trained visual encoder $\phi_{\text{vis}}(\cdot)$ and a frozen pre-trained text encoder $\phi_{\text{txt}}(\cdot)$. Each client k holds a private dataset $\mathcal{D}^{(k)}$. $(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)}) \in \mathcal{D}^{(k)}$, where $\mathbf{x}_i^{(k)}$ is an input image and $\mathbf{y}_i^{(k)} \in \{1, \ldots, C\}$ is the corresponding class label. The full pipeline is illustrated in Figure 2.

3.1 CLIENT SIDE: DUAL-SPACE SUFFICIENT STATISTICS

Without any local training, each client k accumulates per-class sufficient statistics in two complementary feature spaces. One is a visual space produced by extracting sample $\mathbf{x}_i^{(k)} \in \mathcal{D}^{(k)}$ with a frozen pre-trained vision encoder $\phi_{\text{vis}}(\cdot)$. And the other is a low-dimensional text space by projecting the visual space into a class anchor space generated by a frozen pre-trained text encoder $\phi_{\text{txt}}(\cdot)$.

Visual Space ($\mathbb{R}^D$). For each sample $\mathbf{x}_i^{(k)}$, the client extracts an L2-normalized visual feature:

$$\mathbf{f}_i^{\text{vis}} = \phi_{\text{vis}}(\mathbf{x}_i^{(k)}) / \|\phi_{\text{vis}}(\mathbf{x}_i^{(k)})\|_2 \in \mathbb{R}^D. \tag{1}$$

$\mathbf{f}_i^{\text{vis}}$ provides a high-dimensional, general-purpose descriptor that captures fine-grained per-sample variation. For every class c present in the local dataset, the client k then accumulates three per-class

sufficient statistics:

$$\mathbf{N}_c^{\text{vis},(k)} = |\{i : \mathbf{y}_i^{(k)} = c\}|, \quad \mathbf{A}_c^{\text{vis},(k)} = \sum_{i:\mathbf{y}_i^{(k)}=c} \mathbf{f}_i^{\text{vis}}, \quad \mathbf{S}_c^{\text{vis},(k)} = \sum_{i:\mathbf{y}_i^{(k)}=c} \mathbf{f}_i^{\text{vis}} (\mathbf{f}_i^{\text{vis}})^\top, \tag{2}$$

where $\mathbf{N}_c^{\text{vis},(k)}$ is the sample count, $\mathbf{A}_c^{\text{vis},(k)}$ is the feature sum and $\mathbf{S}_c^{\text{vis},(k)}$ is the scatter matrix. Together, these quantities form a complete set of sufficient statistics for Gaussian parameter estimation in the D-dimensional visual space.

Text Space ($\mathbb{R}^C$). We construct the *class anchors* $\{\mathbf{a}_c\}_{c=1}^C \subset \mathbb{R}^D$ once, prior to any statistics computation. Each class name is wrapped in a prompt template (e.g., "a photo of a {class}"), passed through the $\phi_{\text{txt}}(\cdot)$ and L2-normalized, yielding one unit vector $\mathbf{a}_c \in \mathbb{R}^D$ per class. Since the class names are public, every client and the server can compute the same set of anchors independently from the ϕ_{txt}, so no anchor transmission is required. Each client then projects its visual features onto the C class anchors via inner products:

$$\mathbf{f}_i^{\text{txt}} = [\mathbf{a}_1, \ldots, \mathbf{a}_C]^\top \mathbf{f}_i^{\text{vis}} \in \mathbb{R}^C, \tag{3}$$

where the c-th component $\mathbf{a}_c^\top \mathbf{f}_i^{\text{vis}}$ is the cosine similarity between the sample's visual feature and the text anchor of class c. The result is a compact, semantically structured representation in which every dimension explicitly corresponds to one class. Following the same form as Equation (2) with $\mathbf{f}_i^{\text{vis}}$ replaced by $\mathbf{f}_i^{\text{txt}}$, the client accumulates per-class sufficient statistics $\{\mathbf{N}_c^{\text{txt},(k)}, \mathbf{A}_c^{\text{txt},(k)}, \mathbf{S}_c^{\text{txt},(k)}\}$ in the C-dimensional text space.

Upload. Each client k only transmits the two sets of sufficient statistics to the server, and no raw data, features, prototypes, model parameters or gradients are uploaded. Exploiting the symmetry of $\mathbf{S}$, the total per-client communication cost is $C \times [1 + D + D(D+1)/2]$ for the visual space plus $C \times [1 + C + C(C+1)/2]$ for the text space.

3.2 Server Side: Training-Free Classifier Construction

After receiving $\{\mathbf{N}_c^{\text{vis},(k)}, \mathbf{A}_c^{\text{vis},(k)}, \mathbf{S}_c^{\text{vis},(k)}\}_{c=1}^{C^{(k)} \in C}$ and $\{\mathbf{N}_c^{\text{txt},(k)}, \mathbf{A}_c^{\text{txt},(k)}, \mathbf{S}_c^{\text{txt},(k)}\}_{c=1}^{C^{(k)} \in C}$ from all K clients, the server constructs the classifier through a four-step training-free pipeline: (i) partition-invariant aggregation across clients, (ii) Oracle Approximating Shrinkage (OAS) for numerical stability, (iii) Structural Covariance Transfer (SCT) for structure-preserving regularization, and (iv) dual-space QDA/LDA construction and fusion.

3.2.1 Partition-Invariant Aggregation

The server first aggregates client statistics for both spaces $s \in \{\text{vis}, \text{txt}\}$ by summation:

$$\mathbf{N}_c^s = \sum_{k=1}^K \mathbf{N}_c^{s,(k)}, \quad \mathbf{A}_c^s = \sum_{k=1}^K \mathbf{A}_c^{s,(k)}, \quad \mathbf{S}_c^s = \sum_{k=1}^K \mathbf{S}_c^{s,(k)}. \tag{4}$$

From Equation (4), we can obtain $\{\mathbf{N}_c^{\text{vis}}, \mathbf{A}_c^{\text{vis}}, \mathbf{S}_c^{\text{vis}}\}$ and $\{\mathbf{N}_c^{\text{txt}}, \mathbf{A}_c^{\text{txt}}, \mathbf{S}_c^{\text{txt}}\}$. From these aggregated statistics, the per-class Gaussian parameters are obtained in closed form for each space:

$$\boldsymbol{\mu}_c^s = \frac{\mathbf{A}_c^s}{\mathbf{N}_c^s}, \quad \boldsymbol{\Sigma}_c^s = \frac{\mathbf{S}_c^s - \mathbf{A}_c^s (\mathbf{A}_c^s)^\top / \mathbf{N}_c^s}{\mathbf{N}_c^s - 1}, \quad \pi_c^s = \frac{\mathbf{N}_c^s}{\sum_{j=1}^C \mathbf{N}_j^s}, \tag{5}$$

where $\boldsymbol{\mu}_c^{\text{vis}}$ and $\boldsymbol{\mu}_c^{\text{txt}}$ are the class means, $\boldsymbol{\Sigma}_c^{\text{vis}}$ and $\boldsymbol{\Sigma}_c^{\text{txt}}$ are the sample covariances, and π_c^{vis} and π_c^{txt} are the class priors. Because each aggregated statistic is a sum over per-sample contributions, the result does not depend on how samples are distributed across clients. TACO is therefore *partition-invariant*, as formally established in Appendix B. As a consequence, its accuracy is unaffected by the Dirichlet concentration α or the number of clients K.

3.2.2 OAS Covariance Shrinkage

When the number of per-class samples $\mathbf{N}_c^{\text{vis}}$ is far smaller than the visual feature dimension D, the sample covariance $\boldsymbol{\Sigma}_c^{\text{vis}}$ is severely underdetermined. Following (Chen et al., 2010; Turazza et al.,

2026), we first apply OAS:

$$\boldsymbol{\Sigma}_c^{\text{vis, OAS}} = (1-\rho_c)\,\boldsymbol{\Sigma}_c^{\text{vis}} + \rho_c \cdot \frac{\operatorname{tr}(\boldsymbol{\Sigma}_c^{\text{vis}})}{D}\,\mathbf{I}, \tag{6}$$

where $\operatorname{tr}(\cdot)$ denotes the matrix trace, $\mathbf{I}$ is the $D \times D$ identity matrix and the shrinkage intensity $\rho_c \in [0,1]$ is computed analytically from $\boldsymbol{\Sigma}_c^{\text{vis}}$ and $\mathbf{N}_c^{\text{vis}}$ via the OAS formula. OAS interpolates the sample covariance toward a scaled identity to stabilize estimation, but the isotropic target carries no information about class structure.

3.2.3 Structural Covariance Transfer via Cross-Class Regularization

OAS stabilizes each class covariance by shrinking it toward a scaled identity, which discards directional structure shared across classes. We therefore propose *Structural Covariance Transfer* (SCT) as a second regularization step. OAS first produces a stable covariance estimate for every class. SCT then forms its target by averaging the stabilized estimates of the other classes and mixes this target with the class's own OAS estimate.

Step 1: Cross-Class Covariance. For each class c, we compute the uniform average of all other class covariances as

$$\boldsymbol{\Sigma}_c^{\text{vis, cross}} = \frac{1}{C-1}\sum_{j \neq c} \boldsymbol{\Sigma}_j^{\text{vis, OAS}}. \tag{7}$$

Step 2: Adaptive Mixing. The cross-class covariance is then blended with the class's own OAS estimate,

$$(\boldsymbol{\Sigma}_c^{\text{vis}})' = (1-\lambda_c)\,\boldsymbol{\Sigma}_c^{\text{vis, OAS}} + \lambda_c\,\boldsymbol{\Sigma}_c^{\text{vis, cross}}, \quad \lambda_c = \frac{\gamma}{\mathbf{N}_c^{\text{vis}} + \gamma}, \tag{8}$$

where γ controls the regularization strength. Since λ_c decreases with $\mathbf{N}_c^{\text{vis}}$, sample-poor classes place more weight on the cross-class target. Appendix C shows that this target is closer to the reference covariance for that class than the OAS target for nearly all evaluated classes, even when both targets are constructed from few samples. We apply SCT only in the visual space. On the main benchmarks, $\mathbf{N}_c^{\text{txt}} > C$, so the text-space covariance is full rank and left unregularized.

3.2.4 Dual-Space QDA/LDA Fusion

QDA Branch ($C > 20$). Using the SCT-regularized visual covariance $(\boldsymbol{\Sigma}_c^{\text{vis}})'$ from Equation (8) and the raw text covariance $\boldsymbol{\Sigma}_c^{\text{txt}}$, we build a QDA classifier in each space. For a test sample with features $\mathbf{f}^{\text{vis}}$ and $\mathbf{f}^{\text{txt}}$, the per-space log-posteriors are as follows:

$$\begin{aligned}
\ell_c^{\text{QDA, vis}}(\mathbf{f}^{\text{vis}}) &= \underbrace{-\tfrac{1}{2}(\mathbf{f}^{\text{vis}} - \boldsymbol{\mu}_c^{\text{vis}})^\top((\boldsymbol{\Sigma}_c^{\text{vis}})')^{-1}(\mathbf{f}^{\text{vis}} - \boldsymbol{\mu}_c^{\text{vis}})}_{\text{Mahalanobis distance}} \underbrace{-\tfrac{1}{2}\log|(\boldsymbol{\Sigma}_c^{\text{vis}})'|}_{\text{complexity penalty}} + \underbrace{\log \pi_c^{\text{vis}}}_{\text{prior}}, \\
\ell_c^{\text{QDA, txt}}(\mathbf{f}^{\text{txt}}) &= \underbrace{-\tfrac{1}{2}(\mathbf{f}^{\text{txt}} - \boldsymbol{\mu}_c^{\text{txt}})^\top(\boldsymbol{\Sigma}_c^{\text{txt}})^{-1}(\mathbf{f}^{\text{txt}} - \boldsymbol{\mu}_c^{\text{txt}})}_{\text{Mahalanobis distance}} \underbrace{-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_c^{\text{txt}}|}_{\text{complexity penalty}} + \underbrace{\log \pi_c^{\text{txt}}}_{\text{prior}}.
\end{aligned} \tag{9}$$

Both branches share the Bayes-optimal quadratic form but operate on complementary geometries. The D-dimensional visual space retains fine-grained per-sample variation, whereas the C-dimensional text space encodes a semantic coordinate in which each dimension is directly tied to one class. Fusing their posteriors therefore integrates discriminative cues that neither space captures on its own.

LDA Branch ($C \leq 20$). With few classes, the variance of C separate covariance estimates outweighs the class-specific structure they retain, and pooling is preferable. We collapse QDA into LDA with a single shared covariance:

$$\boldsymbol{\Sigma}^{\text{pool, vis}} = \sum_{c=1}^{C} \frac{\mathbf{N}_c^{\text{vis}} - 1}{\sum_{j=1}^{C}(\mathbf{N}_j^{\text{vis}} - 1)}\,\boldsymbol{\Sigma}_c^{\text{vis, OAS}}. \tag{10}$$

Substituting the shared $\boldsymbol{\Sigma}^{\text{pool, vis}}$ into Equation (9) and dropping terms that are constant across c reduces the log-posterior to a linear discriminant:

$$\ell_c^{\text{LDA, vis}}(\mathbf{f}^{\text{vis}}) = (\mathbf{f}^{\text{vis}})^\top(\boldsymbol{\Sigma}^{\text{pool, vis}})^{-1}\boldsymbol{\mu}_c^{\text{vis}} - \tfrac{1}{2}(\boldsymbol{\mu}_c^{\text{vis}})^\top(\boldsymbol{\Sigma}^{\text{pool, vis}})^{-1}\boldsymbol{\mu}_c^{\text{vis}} + \log \pi_c^{\text{vis}}. \tag{11}$$

Because $\mathbf{\Sigma}^{\text{pool, vis}}$ already averages class covariances, SCT provides no additional benefit in this regime. The low-dimensional text space with $C/D \approx 2\%$ is likewise too narrow to carry complementary signal, so we rely on the visual branch alone.

Log-Posterior Fusion. We define the fused per-class log-posterior as:

$$\ell_c'(\mathbf{f}^{\text{vis}}, \mathbf{f}^{\text{txt}}) = \begin{cases} (1-\alpha_f)\,\ell_c^{\text{QDA, vis}}(\mathbf{f}^{\text{vis}}) + \alpha_f\,\ell_c^{\text{QDA, txt}}(\mathbf{f}^{\text{txt}}), & C > 20, \\ \ell_c^{\text{LDA, vis}}(\mathbf{f}^{\text{vis}}), & C \le 20. \end{cases} \tag{12}$$

And predict $\hat{\mathbf{y}} = \arg\max_c \ell_c'(\mathbf{f}^{\text{vis}}, \mathbf{f}^{\text{txt}})$, where $\alpha_f \in (0,1)$ is a fixed fusion weight shared across all datasets with $C > 20$. The sensitivity of α_f is provided in Appendix E.

3.3 Confidence-Weighted Test-Time Augmentation (CTTA)

A single test view carries per-sample feature noise from the frozen encoder, so we aggregate predictions across multiple test-time augmentations. For each test sample, we generate V augmented views with features $\{(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}})\}_{v=1}^{V}$, and evaluate the fused log-posterior $\ell_c'(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}})$ at every view via Equation (12). Stacking across classes yields the per-view vector $\boldsymbol{\ell}_v' = \big(\ell_1'(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}}), \ldots, \ell_C'(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}})\big) \in \mathbb{R}^C$. We then aggregate these vectors through confidence-weighted averaging:

$$\ell_c = \sum_{v=1}^{V} w_v \cdot \ell_c'(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}}),\ w_v = \frac{\exp(\text{conf}_v)}{\sum_{v'} \exp(\text{conf}_{v'})},\ \text{conf}_v = \max_c \text{softmax}(\boldsymbol{\ell}_v'(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}})), \tag{13}$$

where conf_v is the maximum predicted softmax probability for view v. Higher-confidence views receive higher weight, filtering out noisy or ambiguous augmentations before final prediction.

4 Experiments

4.1 Setup

Datasets. We evaluate on three image classification benchmarks: CIFAR-10/100 (Krizhevsky, 2009) (10/100 classes, 32×32) and Tiny-ImageNet (Le & Yang, 2015) (200 classes, 64×64). Appendix H extends the evaluation to six NLP and seven vision benchmarks.

Federated setting. Data is distributed across K=5 clients using Dirichlet sampling with concentration $\alpha \in \{0.05, 0.1, 0.3, 0.5\}$. Smaller α corresponds to stronger non-IID heterogeneity.

Baselines. We compare two categories of baselines: (i) *train-then-merge*: O-FedAvg (McMahan et al., 2017), FedBEns (Talpini et al., 2025), MA-Echo (Su et al., 2023), DENSE (Zhang et al., 2022), Co-Boosting (Dai et al., 2024), IntactOFL (Zeng et al., 2024), FuseFL (Tang et al., 2024) and FAFI (Zeng et al., 2025), and (ii) *frozen-encoder methods*: FALCON (Liu et al., 2026), FedCGS (Guan et al., 2025), Fed3R (Fanì et al., 2024), AFL (He et al., 2025), DeepAFL (Tang et al., 2026), SAFLe (Bacellar et al., 2026), GH-NB/LDA/QDA/Proto-Hyper (Turazza et al., 2026) and FedCOF (Goswami et al., 2025). Table 1 specifies the evaluated backbone configurations. Appendix F describes each baseline.

Backbone. Statistics-based methods use one of two frozen pre-trained visual encoders, CLIP ViT-B/32 and ResNet-18 (He et al., 2016) pre-trained on ImageNet-1K, both with D=512. ResNet-18 is unimodal, so we pair it with the frozen CLIP text encoder for class-name anchors, shared across clients. Train-then-merge baselines train ResNet-18 from scratch on client data, following Zeng et al. (2025).

See Appendix D for implementation details and Appendix E for sensitivity analysis.

4.2 Main Results

Table 1 reports accuracy under four Dirichlet concentrations. FedCOF and FALCON are partition-dependent, since the former estimates covariances from client-level means and the latter fits a feature generator per client.

TACO outperforms or matches prior OFL methods. Under CLIP ViT-B/32, TACO in its default configuration reaches 92.06%/76.69%/71.00% on CIFAR-10/CIFAR-100/Tiny-ImageNet and

exceeds every prior method on CIFAR-100 and Tiny-ImageNet, where per-class estimation is hardest, by $+1.01\%/+1.37\%$ over DeepAFL and $+2.21\%/+2.60\%$ over SAFLe, while matching SAFLe on CIFAR-10. Against FAFI at α=0.5, its most favorable setting, TACO improves by $+3.32\%/+15.62\%/+14.52\%$ without any local training.

TACO generalizes to non-CLIP visual encoders. With a frozen ResNet-18 paired with the CLIP text encoder, TACO in the same configuration outperforms DeepAFL by $+0.05\%/+3.28\%/+3.42\%$ and SAFLe by $+0.13\%/+2.92\%/+3.84\%$ on CIFAR-10/CIFAR-100/Tiny-ImageNet. SCT also improves on OAS with ResNet-18 features (Table 7), so the structure it exploits is not specific to vision-language pretraining. Appendix J reports further encoders.

TACO is not saturated at its default configuration. The results above use hyperparameters selected on CIFAR-100 and then held fixed across datasets and backbones. Adapting the client-side preprocessing, the covariance regularization, and the fusion weight to each setting, at the same upload size, raises TACO to $93.39\%/78.65\%/71.48\%$ under CLIP ViT-B/32 and to $89.15\%/73.97\%/67.21\%$ under ResNet-18. We report this variant as TACO-Tune and describe it in Appendix D.

Table 1: **Accuracy under heterogeneous partitions**. Test accuracy (%) under Dirichlet non-IID with K=5 clients. † marks partition-invariant methods, which are identical across all α. Best per backbone in **bold**, second best underlined, and shading marks TACO.

	CIFAR-10				CIFAR-100				Tiny-ImageNet			
Method	**0.05**	**0.1**	**0.3**	**0.5**	**0.05**	**0.1**	**0.3**	**0.5**	**0.05**	**0.1**	**0.3**	**0.5**
Train-then-merge OFL (ResNet-18 trained from scratch)												
O-FedAvg (AISTATS'17)	12.13	17.43	28.07	35.42	4.77	6.45	10.67	12.13	5.67	8.31	13.61	13.71
DENSE (NeurIPS'22)	38.37	50.26	59.76	62.19	18.37	32.03	37.33	38.84	18.77	22.25	28.14	32.34
MA-Echo (NN'23)	36.77	51.23	60.14	64.21	19.54	29.11	37.77	41.94	15.46	22.23	23.46	28.21
Co-Boosting (ICLR'24)	39.20	58.49	67.21	70.24	20.19	27.59	39.30	42.67	19.00	21.90	29.24	30.78
IntactOFL (MM'24)	48.22	61.13	70.21	79.93	27.99	39.15	41.86	46.78	20.45	28.43	30.15	35.09
FuseFL (NeurIPS'24)	54.42	73.79	84.58	84.34	29.12	36.86	45.12	49.30	22.15	29.28	33.04	34.34
FedBEns (ICML'25)	17.38	15.61	16.40	30.31	5.81	10.35	13.73	15.77	5.70	4.55	8.74	8.40
FAFI (ICML'25)	71.84	77.83	84.76	88.74	31.02	45.48	56.65	61.07	36.96	43.62	53.32	56.48
Frozen ResNet-18 pre-trained on ImageNet-1K												
FedCGS† (AAAI'25)	63.96	63.96	63.96	63.96	39.96	39.96	39.96	39.96	37.54	37.54	37.54	37.54
AFL† (CVPR'25)	80.75	80.75	80.75	80.75	58.56	58.56	58.56	58.56	54.67	54.67	54.67	54.67
FALCON (AAAI'26)	81.97	82.01	85.39	86.45	60.18	61.64	64.10	65.52	57.63	59.45	62.21	63.13
GH-NB† (ICLR'26)	78.84	78.84	78.84	78.84	55.51	55.51	55.51	55.51	55.67	55.67	55.67	55.67
GH-LDA† (ICLR'26)	86.05	86.05	86.05	86.05	63.92	63.92	63.92	63.92	61.06	61.06	61.06	61.06
GH-QDA† (ICLR'26)	84.40	84.40	84.40	84.40	66.52	66.52	66.52	66.52	59.01	59.01	59.01	59.01
GH-Proto-Hyper† (ICLR'26)	85.74	85.74	85.74	85.74	64.05	64.05	64.05	64.05	62.15	62.15	62.15	62.15
DeepAFL† (ICLR'26)	86.43	86.43	86.43	86.43	66.98	66.98	66.98	66.98	62.35	62.35	62.35	62.35
SAFLe† (CVPR'26)	86.35	86.35	86.35	86.35	67.34	67.34	67.34	67.34	61.93	61.93	61.93	61.93
TACO (Ours)†	86.48	86.48	86.48	86.48	70.26	70.26	70.26	70.26	65.77	65.77	65.77	65.77
TACO-Tune (Ours)†	**89.15**	**89.15**	**89.15**	**89.15**	**73.97**	**73.97**	**73.97**	**73.97**	**67.21**	**67.21**	**67.21**	**67.21**
Frozen CLIP ViT-B/32 pre-trained on 400M image-text pairs												
Fed3R† (ICML'24)	91.99	91.99	91.99	91.99	71.99	71.99	71.99	71.99	66.39	66.39	66.39	66.39
FedCOF (NeurIPS'25)	86.81	87.34	87.73	87.54	67.50	67.95	68.57	69.01	63.50	64.09	64.87	64.88
AFL† (CVPR'25)	91.91	91.91	91.91	91.91	70.71	70.71	70.71	70.71	64.80	64.80	64.80	64.80
FALCON (AAAI'26)	90.14	90.01	91.88	91.96	69.70	70.93	72.61	73.87	63.42	67.12	68.92	69.57
GH-LDA† (ICLR'26)	91.10	91.10	91.10	91.10	70.11	70.11	70.11	70.11	64.23	64.23	64.23	64.23
GH-QDA† (ICLR'26)	89.52	89.52	89.52	89.52	72.97	72.97	72.97	72.97	65.67	65.67	65.67	65.67
GH-Proto-Hyper† (ICLR'26)	91.54	91.54	91.54	91.54	73.94	73.94	73.94	73.94	69.29	69.29	69.29	69.29
DeepAFL† (ICLR'26)	91.99	91.99	91.99	91.99	75.68	75.68	75.68	75.68	69.63	69.63	69.63	69.63
SAFLe† (CVPR'26)	92.09	92.09	92.09	92.09	74.48	74.48	74.48	74.48	68.40	68.40	68.40	68.40
TACO (Ours)†	92.06	92.06	92.06	92.06	76.69	76.69	76.69	76.69	71.00	71.00	71.00	71.00
TACO-Tune (Ours)†	**93.39**	**93.39**	**93.39**	**93.39**	**78.65**	**78.65**	**78.65**	**78.65**	**71.48**	**71.48**	**71.48**	**71.48**

Robustness to data heterogeneity. Appendix Figure 3 shows that *train-then-merge* methods degrade sharply as α decreases, while the partition-invariant statistics-based methods remain flat. FAFI drops from 88.74% to 71.84% on CIFAR-10 (-16.90%), and at the extreme α=0.05, reaches only 31.02%

on CIFAR-100, over 40% below TACO at 76.69%. FALCON shares the frozen encoder yet still drops from 86.45% to 81.97% on CIFAR-10 with ResNet-18 as α decreases from 0.5 to 0.05, because its generators inherit each client's label skew. Further results under label noise and long-tailed class counts are reported in Appendix K.

4.3 Ablation Study

SCT is the largest contributor, and its benefit scales with estimation difficulty. Adding SCT to OAS-regularized QDA yields +0.22%/+3.79%/+5.57% on CIFAR-10/100/Tiny-ImageNet (Table 2), matching covariance-estimation difficulty. CIFAR-10 has 5000 samples per class with $\mathbf{N}_c \gg D$, so estimation is already reliable, whereas Tiny-ImageNet has 500 samples per class in D=512 dimensions, where OAS's isotropic target leaves bias that SCT corrects.

The preferred covariance model differs across datasets. LDA reaches 91.60% on CIFAR-10, above QDA+OAS+SCT at 90.36%, whereas QDA+OAS+SCT exceeds LDA on CIFAR-100 and Tiny-ImageNet (75.38%/69.24% versus 70.60%/65.68%). Across the LDA rows, OAS or SCT changes accuracy by at most 0.23%, so the SCT gains are concentrated in the per-class QDA setting.

SCT and text fusion each remain effective with or without the other. SCT raises accuracy by 3.79% and 5.57% on CIFAR-100 and Tiny-ImageNet without fusion and by 3.99% and 5.14% with it, while text fusion adds 0.26% and 0.57% to QDA+OAS and 0.46% and 0.14% to QDA+OAS+SCT.

CTTA improves accuracy at inference. Adding CTTA to the full configuration improves accuracy by 0.46%/0.85%/1.62% on CIFAR-10/100/Tiny-ImageNet without changing the aggregated statistics or the server-side classifier.

Table 2: **Component ablation**. All results (%) are partition-invariant. †Cosine: zero-shot nearest-anchor classification. *Adaptive: LDA without fusion on CIFAR-10, QDA with α_f=0.25 otherwise. Arrows show within-block changes from the preceding row, with gains in red and others in gray. **Bold**/underline mark the best/second-best results, and shading marks full TACO.

Configuration	CIFAR-10	CIFAR-100	Tiny-ImageNet
Cosine baseline†	83.54	58.21	57.18
QDA-based (visual space)			
No shrinkage	90.13	67.80	58.99
OAS	90.14 ↑0.01	71.59 ↑3.79	63.67 ↑4.68
OAS + SCT	90.36 ↑0.22	75.38 ↑3.79	69.24 ↑5.57
LDA-based (visual space)			
No shrinkage	91.60	70.60	65.68
OAS	91.60 →0.00	70.37 ↓0.23	65.47 ↓0.21
OAS + SCT	91.60 →0.00	70.35 ↓0.02	65.48 ↑0.01
QDA-based (text space)			
No shrinkage	88.15	71.47	65.71
TACO (adaptive, visual + text)*			
OAS + fusion*	91.60	71.85	64.24
OAS + SCT + fusion*	91.60 →0.00	75.84 ↑3.99	69.38 ↑5.14
OAS + SCT + fusion* + CTTA	**92.06** ↑0.46	**76.69** ↑0.85	**71.00** ↑1.62

5 Conclusion

We presented TACO, a training-free OFL framework applicable to any frozen visual encoder. TACO improves statistical estimation along three complementary axes: (i) *Structural Covariance Transfer* replaces OAS's isotropic shrinkage target with a cross-class covariance average, letting sample-poor classes borrow directional structure from sample-rich ones, (ii) *dual-space fusion* adds a text-anchor view with complementary decision geometry, (iii) *confidence-weighted test-time augmentation* suppresses per-sample feature noise across augmented views. TACO outperforms or matches all compared SOTA OFL methods on CIFAR-10, CIFAR-100 and Tiny-ImageNet, reaching 92.06%/76.69%/71.00% accuracy, while requiring no gradient-based training and remaining strictly partition-invariant.

AI USE STATEMENT

In this work, We used generative AI tools solely to polish author-written text and improve the readability and fluency of the manuscript. These tools were not used for any other aspect of this work, including research ideation, methodology development, theoretical analysis, code implementation, data preparation, experimental design or execution, or interpretation of results. The authors reviewed all AI-assisted edits to ensure that they preserved the intended meaning and technical accuracy of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The client and server procedures, sufficient statistics, SCT estimator, dual-space fusion rule, and test-time inference are specified in Sections 3 and 4 and Algorithm 1. Dataset partitions, preprocessing, hyperparameters, implementation details, additional ablations, and sensitivity analyses are documented in the appendix.

REFERENCES

Youssef Allouah, Akash Dhasade, Rachid Guerraoui, Nirupam Gupta, Anne-Marie Kermarrec, Rafael Pinot, Rafael Pires, and Rishi Sharma. Revisiting ensembling in one-shot federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 68500–68527, 2024. doi: 10.52202/079017-2188.

Alan T. L. Bacellar, Mustafa Munir, Felipe M. G. França, Priscila M. V. Lima, Radu Marculescu, and Lizy K. John. Single-round scalable analytic federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39445–39454, 2026.

Mahdi Beitollahi, Alex Bie, Sobhan Hemati, Leo Maxime Brunswic, Xu Li, Xi Chen, and Guojun Zhang. Foundation models meet federated learning: A one-shot feature-sharing method with privacy and performance guarantees. *Transactions on Machine Learning Research (TMLR)*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=55593xywWG.

Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004. doi: 10.3150/bj/1106314847.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI (NLP4ConvAI)*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, and Volker Tresp. FedBiP: Heterogeneous one-shot federated learning with personalized latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30440–30450, 2025. doi: 10.1109/CVPR52734.2025.02834.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing (TSP)*, 58(10):5016–5029, 2010. doi: 10.1109/TSP.2010.2053029.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Rong Dai, Yonggang Zhang, Ang Li, Tongliang Liu, Xun Yang, and Bo Han. Enhancing one-shot federated learning through data and ensemble co-boosting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=tm8s3696Ox.

Khalil Elkhalil, Abla Kammoun, Romain Couillet, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini. A large dimensional study of regularized discriminant analysis. *IEEE Transactions on Signal Processing (TSP)*, 68:2464–2479, 2020. doi: 10.1109/TSP.2020.2984160.

Eros Fanì, Raffaello Camoriano, Barbara Caputo, and Marco Ciccone. Accelerating heterogeneous federated learning with closed-form classifiers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13029–13048. PMLR, 2024. URL https://proceedings.mlr.press/v235/fani-24a.html.

Dipam Goswami, Simone Magistri, Kai Wang, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost van de Weijer. Covariances for free: Exploiting mean distributions for training-free federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 65081–65115, 2025. doi: 10.52202/085713-2179. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5df817c5dd95293ebf6d1583303a8c73-Abstract-Conference.html.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007. URL https://authors.library.caltech.edu/records/5sv1j-ytw97.

Zenghao Guan, Yucan Zhou, and Xiaoyan Gu. Capture global feature statistics for one-shot federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 39(16):16942–16950, 2025. doi: 10.1609/aaai.v39i16.33862. URL https://ojs.aaai.org/index.php/AAAI/article/view/33862.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, and Huiping Zhuang. AFL: A single-round analytic approach for federated learning with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4988–4998, 2025. doi: 10.1109/CVPR52734.2025.00470.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedFisher: Leveraging Fisher information for one-shot federated learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1612–1620. PMLR, 2024. URL https://proceedings.mlr.press/v238/jhunjhunwala24a.html.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html.

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

Ken Lang. NewsWeeder: Learning to filter netnews. In *Machine Learning Proceedings 1995 (ICML)*, pp. 331–339. Morgan Kaufmann, 1995. doi: 10.1016/B978-1-55860-377-6.50048-7.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1311–1316. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131/.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N course project report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis (JMVA)*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Gwen Legate, Nicolas Bernier, Lucas Page-Caccia, Edouard Oyallon, and Eugene Belilovsky. Guiding the last layer in federated learning with pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 69832–69848, 2023. doi: 10.52202/075280-3061. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcc0ac74ac8b95dc1939804acce0317d-Abstract-Conference.html.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10713–10722, 2021. doi: 10.1109/CVPR46437.2021.01057. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Model-Contrastive_Federated_Learning_CVPR_2021_paper.html.

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

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 2351–2363, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/18df51b97ccd68128e994804f3eccc87-Abstract.html.

Shudong Liu, Hanwen Zhang, Xiuling Wang, Yuesheng Zhu, and Guibo Luo. Feature-aware one-shot federated learning via hierarchical token sequences. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 40(28):23819–23827, 2026. doi: 10.1609/aaai.v40i28.39557.

Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. FedLPA: One-shot federated learning with layer-wise posterior aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 81510–81548, 2024. doi: 10.52202/079017-2591.

Xiang Liu, Zhenheng Tang, Xia Li, Yijun Song, Sijie Ji, Zemin Liu, Bo Han, Linshan Jiang, and Jialin Li. One-shot federated learning methods: A practical guide. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 10573–10581, 2025. doi: 10.24963/ijcai.2025/1174. URL https://www.ijcai.org/proceedings/2025/1174. Survey Track.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction (IWSDS)*, volume 714 of *Lecture Notes in Electrical Engineering*, pp. 165–183. Springer, 2021. doi: 10.1007/978-981-15-9323-9_15.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of*

the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729. IEEE, 2008. doi: 10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505. IEEE, 2012. doi: 10.1109/CVPR.2012.6248092.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology (SAGMB)*, 4(1), 2005. doi: 10.2202/1544-6115.1175. Article 32.

Rui Song, Dai Liu, Dave Zhenyu Chen, Andreas Festag, Carsten Trinitis, Martin Schulz, and Alois Knoll. Federated learning via decentralized dataset distillation in resource-constrained edge environments. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE, 2023. doi: 10.1109/IJCNN54540.2023.10191879. URL https://ieeexplore.ieee.org/document/10191879.

Shangchao Su, Bin Li, and Xiangyang Xue. One-shot federated learning without server-side training. *Neural Networks*, 164:203–215, 2023. doi: 10.1016/j.neunet.2023.04.035.

Jacopo Talpini, Marco Savi, and Giovanni Neglia. FedBEns: One-shot federated learning based on Bayesian ensemble. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58448–58466. PMLR, 2025. URL https://proceedings.mlr.press/v267/talpini25a.html.

Jianheng Tang, Yajiang Huang, Kejia Fan, Feijiang Han, Jiaxu Li, Jinfeng Xu, Run He, Anfeng Liu, Houbing Herbert Song, Huiping Zhuang, and Yunhuai Liu. DeepAFL: Deep analytic federated learning. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=ve3EzAvMGe.

Zhenheng Tang, Yonggang Zhang, Peijie Dong, Yiu-ming Cheung, Amelie Chi Zhou, Bo Han, and Xiaowen Chu. FuseFL: One-shot federated learning through the lens of causality with progressive model fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 28393–28429, 2024. doi: 10.52202/079017-0891.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

TorchVision maintainers and contributors. TorchVision: PyTorch's computer vision library. https://github.com/pytorch/vision, 2016.

Fabio Turazza, Marco Picone, and Marco Mamei. The Gaussian-Head OFL family: One-shot federated learning from client global statistics. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0da2c78fed88ad2d368994970a430c96-Abstract-Conference.html.

Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL https://arxiv.org/abs/2110.00476.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492. IEEE, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. URL https://arxiv.org/abs/1905.00546.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=JWOiYxMG92s. Oral.

Hui Zeng, Minrui Xu, Tongqing Zhou, Xinyi Wu, Jiawen Kang, Zhiping Cai, and Dusit Niyato. One-shot-but-not-degraded federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 11070–11079, 2024. doi: 10.1145/3664647.3680715.

Hui Zeng, Wenke Huang, Tongqing Zhou, Xinyi Wu, Guancheng Wan, Yingwen Chen, and Zhiping Cai. Does one-shot give the best shot? Mitigating model inconsistency in one-shot federated learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74080–74097. PMLR, 2025. URL https://proceedings.mlr.press/v267/zeng25c.html.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. DENSE: Data-free one-shot federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 21414–21428, 2022. doi: 10.52202/068431-1556.

Li Zhang, Zhongxuan Han, XiaoHua Feng, Jiaming Zhang, Yuyuan Li, Linbo Jiang, Jianan Lin, and Chaochao Chen. TOFA: Training-free one-shot federated adaptation for vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 40(33):28301–28309, 2026. doi: 10.1609/aaai.v40i33.40058. URL https://ojs.aaai.org/index.php/AAAI/article/view/40058.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html.

A TACO ALGORITHM

Algorithm 1 summarizes the full TACO procedure.

Algorithm 1 TACO: Text-Anchored Classification with Structural Covariance Transfer

Require: K: number of clients, C: number of classes, $\{\mathcal{D}^{(k)}\}_{k=1}^{K}$: local private datasets, $\phi_{\text{vis}}(\cdot)/\phi_{\text{txt}}(\cdot)$: frozen pre-trained visual/text encoders, γ: SCT pseudo-count, α_f: dual-space fusion weight, V: number of CTTA augmented views, $s \in \{\text{vis}, \text{txt}\}$.
Ensure: Predicted label $\hat{\mathbf{y}}$ for each test sample $\mathbf{x}$.

Every client and the server build class anchors $\{\mathbf{a}_c\}_{c=1}^{C}$ as L2-normalized $\phi_{\text{txt}}(\cdot)$ embeddings of the class-name prompts
ClientStatistics(k):
for each $(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(k)}) \in \mathcal{D}^{(k)}$ **do**
  $\mathbf{f}_i^{\text{vis}} \leftarrow \phi_{\text{vis}}(\mathbf{x}_i^{(k)})/\|\phi_{\text{vis}}(\mathbf{x}_i^{(k)})\|_2 \in \mathbb{R}^D$ ▷ Eq. (1)
  $\mathbf{f}_i^{\text{txt}} \leftarrow [\mathbf{a}_1, \ldots, \mathbf{a}_C]^\top \mathbf{f}_i^{\text{vis}} \in \mathbb{R}^C$ ▷ Eq. (3)
end for
for each space $s \in \{\text{vis}, \text{txt}\}$ and class c **do**
  $\mathbf{N}_c^{s,(k)} \leftarrow |\{i : \mathbf{y}_i^{(k)} = c\}|,\ \mathbf{A}_c^{s,(k)} \leftarrow \sum_{i:\mathbf{y}_i^{(k)}=c} \mathbf{f}_i^s,\ \mathbf{S}_c^{s,(k)} \leftarrow \sum_{i:\mathbf{y}_i^{(k)}=c} \mathbf{f}_i^s (\mathbf{f}_i^s)^\top$ ▷ Eq. (2)
end for
return $\{\mathbf{N}_c^{s,(k)}, \mathbf{A}_c^{s,(k)}, \mathbf{S}_c^{s,(k)}\}_{c=1}^{C^{(k)} \in C}$ to server
ServerExecution:
for each client $k \in \{1, \ldots, K\}$ **in parallel do**
  $\{\mathbf{N}_c^{s,(k)}, \mathbf{A}_c^{s,(k)}, \mathbf{S}_c^{s,(k)}\} \leftarrow$ **ClientStatistics**(k)
end for
$\mathbf{N}_c^s \leftarrow \sum_{k=1}^{K} \mathbf{N}_c^{s,(k)},\ \mathbf{A}_c^s \leftarrow \sum_{k=1}^{K} \mathbf{A}_c^{s,(k)},\ \mathbf{S}_c^s \leftarrow \sum_{k=1}^{K} \mathbf{S}_c^{s,(k)}$ ▷ Eq. (4)
$\boldsymbol{\mu}_c^s \leftarrow \mathbf{A}_c^s/\mathbf{N}_c^s,\ \boldsymbol{\Sigma}_c^s \leftarrow \left(\mathbf{S}_c^s - \mathbf{A}_c^s(\mathbf{A}_c^s)^\top/\mathbf{N}_c^s\right)/(\mathbf{N}_c^s - 1),\ \pi_c^s \leftarrow \mathbf{N}_c^s/\sum_{j=1}^{C} \mathbf{N}_j^s$ ▷ Eq. (5)
$\boldsymbol{\Sigma}_c^{\text{vis,OAS}} \leftarrow (1-\rho_c)\,\boldsymbol{\Sigma}_c^{\text{vis}} + \rho_c \frac{\text{tr}(\boldsymbol{\Sigma}_c^{\text{vis}})}{D}\,\mathbf{I}$, with ρ_c via OAS (Chen et al., 2010) ▷ Eq. (6)
if $C > 20$ **then**
  $\boldsymbol{\Sigma}_c^{\text{vis,cross}} \leftarrow \frac{1}{C-1} \sum_{j \neq c} \boldsymbol{\Sigma}_j^{\text{vis,OAS}}$ ▷ Eq. (7)
  $\lambda_c \leftarrow \frac{\gamma}{\mathbf{N}_c^{\text{vis}}+\gamma},\ (\boldsymbol{\Sigma}_c^{\text{vis}})' \leftarrow (1-\lambda_c)\,\boldsymbol{\Sigma}_c^{\text{vis,OAS}} + \lambda_c\,\boldsymbol{\Sigma}_c^{\text{vis,cross}}$ ▷ Eq. (8)
  Build the QDA scores $\ell_c^{\text{QDA,vis}}$ and $\ell_c^{\text{QDA,txt}}$ with Eq. (9)
else
  $\boldsymbol{\Sigma}^{\text{pool,vis}} \leftarrow \sum_c \frac{\mathbf{N}_c^{\text{vis}}-1}{\sum_j (\mathbf{N}_j^{\text{vis}}-1)}\,\boldsymbol{\Sigma}_c^{\text{vis,OAS}}$ ▷ Eq. (10)
  Build the visual LDA score with Eq. (11)
  $\alpha_f \leftarrow 0$
end if
return $\hat{\mathbf{y}} \leftarrow$ **Inference**$(\mathbf{x})$ for each test sample $\mathbf{x}$
Inference$(\mathbf{x})$:
Generate V augmented views of $\mathbf{x}$
for $v = 1$ **to** V **do**
  Extract $\mathbf{f}_v^{\text{vis}}$ and $\mathbf{f}_v^{\text{txt}}$ from view v ▷ Eqs. (1) and (3)
  Evaluate $\boldsymbol{\ell}_v^{'}$ with the QDA-fusion or visual-LDA branch in Eq. (12)
  $\text{conf}_v \leftarrow \max_c[\text{softmax}(\boldsymbol{\ell}_v')]_c$
end for
$w_v \leftarrow \frac{\exp(\text{conf}_v)}{\sum_{v'} \exp(\text{conf}_{v'})},\ \ell_c \leftarrow \sum_v w_v\, \ell_c^{'}(\mathbf{f}_v^{\text{vis}}, \mathbf{f}_v^{\text{txt}})$ ▷ Eq. (13)
return $\hat{\mathbf{y}} = \arg\max_c \ell_c$

B PARTITION INVARIANCE: THEORY AND EMPIRICAL VERIFICATION

For a fixed pooled dataset and shared encoders, TACO's closed-form classifier depends only on additive sufficient statistics. Aggregation, covariance regularization, and classifier construction are therefore invariant to the client partition.

Proposition 1 (Partition Invariance). *For any partition $\mathcal{P} = \{\mathcal{D}^{(k)}\}_{k=1}^{K}$ of a fixed dataset $\mathcal{D}$ in which each sample appears on exactly one client, and for each space $s \in \{vis, txt\}$, the aggregated statistics $\{\mathbf{N}_c^s, \mathbf{A}_c^s, \mathbf{S}_c^s\}$, the derived Gaussian parameters $\{\boldsymbol{\mu}_c^s, \boldsymbol{\Sigma}_c^s, \pi_c^s\}$, and therefore TACO's closed-form classifier are identical to those computed centrally from $\mathcal{D}$.*

Proof. Fix a space $s \in \{\text{vis}, \text{txt}\}$. On client k, each statistic is a sum over local samples, for example $\mathbf{A}_c^{s,(k)} = \sum_{i \in \mathcal{D}^{(k)}:\, \mathbf{y}_i = c} \mathbf{f}_i^s$. Addition is commutative and associative, so

$$\sum_k \mathbf{A}_c^{s,(k)} = \sum_k \sum_{i \in \mathcal{D}^{(k)}:\, \mathbf{y}_i = c} \mathbf{f}_i^s = \sum_{i \in \mathcal{D}:\, \mathbf{y}_i = c} \mathbf{f}_i^s.$$

The same identity holds for $\mathbf{N}_c^s$ and $\mathbf{S}_c^s$. The Gaussian parameters, OAS and SCT covariances, and classifier scores are deterministic functions of these aggregated statistics. The final classifier is therefore unchanged when the fixed dataset $\mathcal{D}$ is repartitioned across clients. □

Empirical Verification. Table 3 verifies Proposition 1 on CLIP ViT-B/32. Holding the training set and backbone fixed, we partition the data among $K \in \{5, 10, 20, 50, 100, 5{,}000, 50{,}000\}$ clients at Dirichlet α=0.1 (highly non-IID). Accuracy remains constant at 91.60% / 75.84% / 69.38% on CIFAR-10 / CIFAR-100 / Tiny-ImageNet across all seven values of K. TACO is also flat across $\alpha \in \{0.05, 0.1, 0.3, 0.5\}$ (see Table 1). These results confirm that neither the number of clients nor the degree of label skew affects TACO's output.

Table 3: **Scalability across client counts**. Closed-form TACO accuracy (%) under Dirichlet α=0.1 with CLIP ViT-B/32. Proposition 1 gives the same classifier across all values of K.

K	CIFAR-10	CIFAR-100	Tiny-ImageNet
5	91.60	75.84	69.38
10	91.60	75.84	69.38
20	91.60	75.84	69.38
50	91.60	75.84	69.38
100	91.60	75.84	69.38
5,000	91.60	75.84	69.38
50,000	91.60	75.84	69.38

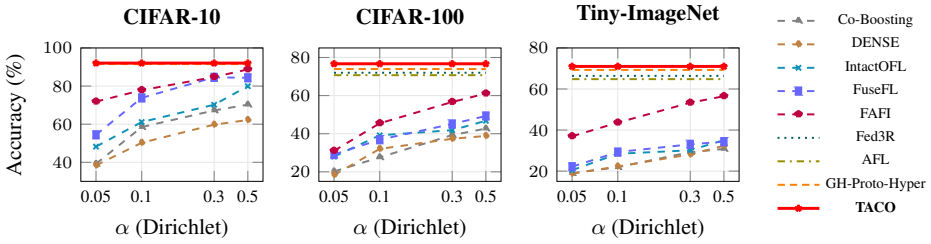

Figure 3: Accuracy vs. Dirichlet concentration α. Dashed lines show *train-then-merge* methods, which degrade under strong non-IID as $\alpha \to 0$. Flat lines show *statistics-based* methods Fed3R, AFL, GH-Proto-Hyper and TACO, which are partition-invariant.

C Why Cross-Class Covariance Is a Better Shrinkage Target

OAS shrinks each class covariance toward a scaled identity. This target assigns the same variance to every direction and contains no directional structure. SCT instead uses the average covariance of the other classes. This choice assumes that different classes share the directions along which their features vary most. We first derive the condition under which the cross-class target gives the more accurate covariance estimate, and what the number of classes contributes to that condition. We then test the assumption. First, we ask whether the main covariance directions are aligned across classes. Second, we ask whether those shared directions carry a substantial part of each class's variance.

Third, we ask whether the cross-class target is closer to the class covariance than the scaled identity. We then use a rotation control to separate directional alignment from eigenvalue magnitudes. We next repeat the distance comparison when both targets are built from few samples. We then vary the samples per class to measure covariance stability and TACO accuracy. Finally, we hold the QDA classifier fixed and compare SCT with alternative covariance estimators.

C.1 Bias and Variance of the Two Shrinkage Targets

Both methods plug a different target into the same estimator, so they can be compared term by term. Write the shrinkage estimator for class c as

$$\hat{\mathbf{\Sigma}}_c(\rho, T) = (1-\rho)\,\hat{\mathbf{\Sigma}}_c + \rho\, T, \qquad \rho \in (0,1), \tag{14}$$

where $\hat{\mathbf{\Sigma}}_c$ is the unbiased sample covariance built from the $\mathbf{N}_c$ samples of class c, so that $\mathbb{E}[\hat{\mathbf{\Sigma}}_c] = \mathbf{\Sigma}_c^*$ for the unknown true covariance $\mathbf{\Sigma}_c^*$. OAS and SCT differ only in the target,

$$\hat{T}_{\mathrm{OAS}} = \frac{\mathrm{tr}(\hat{\mathbf{\Sigma}}_c)}{D}\,\mathbf{I}, \qquad \hat{T}_{\mathrm{SCT}} = \frac{1}{C-1}\sum_{j\neq c}\hat{\mathbf{\Sigma}}_j. \tag{15}$$

Both targets are estimated. Replacing every $\hat{\mathbf{\Sigma}}$ in Equation (15) by the corresponding $\mathbf{\Sigma}^*$ gives the population targets T^*_{OAS} and T^*_{SCT}, for which the two estimates are unbiased. They differ in what they are estimated from. The isotropic target is a function of the samples of class c, while the cross-class target is built from the samples of the other classes and is therefore independent of $\hat{\mathbf{\Sigma}}_c$.

Proposition 2 (Shrinkage bias)**.** *For either target the bias of the estimator in Equation* (14) *equals* $\rho\,(T^* - \mathbf{\Sigma}_c^*)$*. Consequently*

$$\big\|\mathrm{Bias}(\hat{\mathbf{\Sigma}}_c^{\mathrm{SCT}})\big\|_F \le \big\|\mathrm{Bias}(\hat{\mathbf{\Sigma}}_c^{\mathrm{OAS}})\big\|_F \quad\Longleftrightarrow\quad \big\|T^*_{\mathrm{SCT}} - \mathbf{\Sigma}_c^*\big\|_F \le \big\|T^*_{\mathrm{OAS}} - \mathbf{\Sigma}_c^*\big\|_F. \tag{16}$$

Proof. The estimator is affine in $\hat{\mathbf{\Sigma}}_c$ and $\hat{T}$, so its expectation is $(1-\rho)\mathbf{\Sigma}_c^* + \rho T^*$ and its bias has squared Frobenius norm $\rho^2\|T^* - \mathbf{\Sigma}_c^*\|_F^2$. Both estimators use the same positive ρ, so the two biases are ordered exactly as the two target distances. □

Proposition 2 turns a modeling assumption into a measurable one, and Table 4 estimates the two distances with full-sample covariances in place of the population covariances.

Bias is only one part of the estimation error. The cross-class target is itself an average of $C-1$ estimated covariances, so it carries a variance of its own.

Proposition 3 (Shrinkage variance)**.** *Let V_j be the mean squared error $\mathbb{E}\|\hat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j^*\|_F^2$ of the unregularized sample covariance of class j, and $\bar{V}_{-c}$ its average over the classes other than c. Assume the samples of different classes are independent, and write* Var *for* $\mathbb{E}\|\cdot - \mathbb{E}[\cdot]\|_F^2$*. Then*

$$\mathrm{Var}\big(\hat{\mathbf{\Sigma}}_c^{\mathrm{SCT}}\big) = (1-\rho)^2\,V_c + \rho^2\,\frac{\bar{V}_{-c}}{C-1}, \tag{17}$$

$$\mathrm{Var}\big(\hat{\mathbf{\Sigma}}_c^{\mathrm{OAS}}\big) = (1-\rho)^2\,V_c + \rho\,(2-\rho)\,\tau_c^2, \qquad \tau_c^2 = \tfrac{1}{D}\,\mathbb{E}\big[\big(\mathrm{tr}(\hat{\mathbf{\Sigma}}_c - \mathbf{\Sigma}_c^*)\big)^2\big]. \tag{18}$$

Proof. Write $\Delta_c = \hat{\mathbf{\Sigma}}_c - \mathbf{\Sigma}_c^*$ and $\Delta_T = \hat{T} - T^*$, so the centered estimator is $(1-\rho)\Delta_c + \rho\Delta_T$ and

$$\mathrm{Var} = (1-\rho)^2\,\mathbb{E}\|\Delta_c\|_F^2 + 2\rho(1-\rho)\,\mathbb{E}\langle\Delta_c, \Delta_T\rangle + \rho^2\,\mathbb{E}\|\Delta_T\|_F^2. \tag{19}$$

For SCT the target depends only on the classes $j \neq c$, which are independent of the samples of class c, so the middle term factorizes into $\langle\mathbb{E}\Delta_c, \mathbb{E}\Delta_T\rangle = 0$. Independence across classes also removes the cross terms inside $\mathbb{E}\|\Delta_T\|_F^2 = (C-1)^{-2}\sum_{j\neq c} V_j = \bar{V}_{-c}/(C-1)$, which gives Equation (17). For OAS the target fluctuation is $\Delta_T = D^{-1}\,\mathrm{tr}(\Delta_c)\,\mathbf{I}$, hence both $\|\Delta_T\|_F^2$ and $\langle\Delta_c, \Delta_T\rangle$ equal $(\mathrm{tr}\,\Delta_c)^2/D$. Substituting both into Equation (19) and collecting the τ_c^2 terms gives Equation (18). □

The two variances differ in two terms. The middle term vanishes for SCT and stays positive for OAS, because the isotropic target is the trace of the matrix being shrunk. The last term is the variance of the target itself, which comes from one scalar for OAS and from an average of $C-1$ estimated matrices for SCT. Equation (21) states when the closer target is worth this variance.

Corollary 4 (Precision of the cross-class target)**.** *Suppose every class is Gaussian with* $\mathbf{N}$ *samples and shares the value* $\kappa = (\operatorname{tr}\boldsymbol{\Sigma}^*)^2 + \|\boldsymbol{\Sigma}^*\|_F^2$*. Then* $V_j = \kappa/(\mathbf{N}-1)$ *for every class, and the variance contributed by the cross-class target is*

$$\frac{\bar{V}_{-c}}{C-1} = \frac{\kappa}{(C-1)(\mathbf{N}-1)}. \tag{20}$$

The cross-class target is therefore as precise as a covariance estimated from $(C-1)(\mathbf{N}-1)+1$ *samples.*

Proof. For a Gaussian sample with $n = \mathbf{N}-1$ degrees of freedom the entrywise variance is $\operatorname{Var}(\hat{\boldsymbol{\Sigma}}_{ab}) = n^{-1}\big(\boldsymbol{\Sigma}^*_{aa}\boldsymbol{\Sigma}^*_{bb} + (\boldsymbol{\Sigma}^*_{ab})^2\big)$. Summing over all entries gives $V_j = \kappa/(\mathbf{N}-1)$. Substituting into $\bar{V}_{-c}/(C-1)$ gives Equation (20), and matching that value to $\kappa/(n'-1)$ yields $n' = (C-1)(\mathbf{N}-1)+1$. □

Substituting C and $\mathbf{N}$ into Equation (20), the cross-class target reaches the precision of about 4.9×10^4 samples on CIFAR-100 and 9.9×10^4 on Tiny-ImageNet, against 500 for the class covariance itself.

Corollary 5 (When the cross-class target wins)**.** *Under the assumptions of Proposition 3 the SCT estimator has the smaller mean squared error whenever*

$$\big\|T^*_{\mathrm{OAS}} - \boldsymbol{\Sigma}^*_c\big\|_F^2 - \big\|T^*_{\mathrm{SCT}} - \boldsymbol{\Sigma}^*_c\big\|_F^2 \;\geq\; \frac{\bar{V}_{-c}}{C-1}. \tag{21}$$

Proof. Adding the squared bias of Proposition 2 to the variances of Proposition 3 and subtracting one estimator from the other,

$$\mathrm{MSE}_{\mathrm{SCT}} - \mathrm{MSE}_{\mathrm{OAS}} = \rho^2\left[\|T^*_{\mathrm{SCT}} - \boldsymbol{\Sigma}^*_c\|_F^2 - \|T^*_{\mathrm{OAS}} - \boldsymbol{\Sigma}^*_c\|_F^2 + \frac{\bar{V}_{-c}}{C-1}\right] - \rho(2-\rho)\,\tau_c^2.$$

The final term is never positive, so a non-positive bracket is sufficient, which is Equation (21). □

The right side of Condition (21) falls like $1/(C-1)$, while the left side changes with C only through the composition of the cross-class average. SCT gains little on CIFAR-10 for another reason, since 5000 samples per class keep the class's own error V_c small and leave little for either target to remove.

Corollary 6 (Shrinkage weight)**.** *Under the assumptions of Proposition 3 the weight that minimizes* $\mathrm{MSE}(\hat{\boldsymbol{\Sigma}}_c^{\mathrm{SCT}})$ *is*

$$\rho_c^\star = \frac{V_c}{V_c + B_c + \bar{V}_{-c}/(C-1)}, \qquad B_c = \big\|T^*_{\mathrm{SCT}} - \boldsymbol{\Sigma}^*_c\big\|_F^2. \tag{22}$$

With $V_c = \kappa_c/(\mathbf{N}_c-1)$ *as in Corollary 4 this becomes*

$$\rho_c^\star = \frac{\gamma_c}{(\mathbf{N}_c-1)+\gamma_c}, \qquad \gamma_c = \frac{\kappa_c}{B_c + \bar{V}_{-c}/(C-1)}. \tag{23}$$

Proof. By Propositions 2 and 3 the mean squared error is $f(\rho) = \rho^2 B_c + (1-\rho)^2 V_c + \rho^2 \bar{V}_{-c}/(C-1)$, a strictly convex quadratic in ρ. Setting $f'(\rho) = 0$ gives Equation (22), and dividing its numerator and denominator by $B_c + \bar{V}_{-c}/(C-1)$ gives Equation (23). □

Equation (23) has the form of the SCT weight $\lambda_c = \gamma/(\mathbf{N}_c+\gamma)$ in Equation (8). The oracle pseudo-count γ_c is the ratio between the scale of a class's own estimation error and the squared distance from the cross-class target to that class's covariance plus the variance of that target. TACO replaces this class-dependent oracle value by one shared γ, and Appendix E tests how far that value transfers across datasets, encoders and modalities.

These results concern the covariance estimate rather than the classifier. How estimation error turns into misclassification also depends on the discriminant, and when D and $\mathbf{N}_c$ are comparable the error rate of discriminant analysis is not a monotone function of covariance mean squared error (Bickel & Levina, 2004; Elkhalil et al., 2020). The rest of this appendix therefore measures the two target distances of Equation (16) and then the resulting accuracy.

C.2 Shared Directions Explain the Smaller Target Distance

Mean angle measures how closely the leading covariance directions align across classes. These directions capture the largest variations around each class mean. Smaller angles indicate stronger alignment, with random directions giving a reference of about $80°$.

Variance in shared directions measures the fraction of class c's variance captured by the leading directions estimated from the other classes. A larger value means that these directions explain more of class c's variation. For example, a table entry of 0.29 means that the shared directions contain 29% of a class's variance on average. Random directions capture only 0.02-0.04.

We next compare the Frobenius distances of the cross-class and scaled-identity targets to each class's full-sample covariance. A target with unequal eigenvalues could be closer simply because its variance differs across directions, even if those directions are wrong. To separate this possibility from directional alignment, we rotate only the SCT target. Rotation keeps its variance magnitudes but changes the directions assigned to them. We then repeat the same target-distance comparison.

Table 4: **Shared covariance directions and target distance**. Original gives the percentage of classes whose SCT target is closer to the full-sample class covariance than the OAS target. After rotation uses the rotated SCT target, which has the same eigenvalues but different eigenvectors. Parentheses give the mean distance reduction relative to OAS before rotation.

		Shared directions		SCT closer than OAS (% of classes)	
Encoder	**Dataset**	**Mean angle (°) ↓**	**Variance in shared directions ↑**	**Original**	**After rotation**
CLIP ViT-B/32	CIFAR-10	54	0.39	100% (20%)	0%
	CIFAR-100	58	0.40	100% (21%)	0%
	Tiny-ImageNet	64	0.31	100% (13%)	0%
ResNet-18	CIFAR-10	57	0.34	100% (22%)	0%
	CIFAR-100	62	0.33	100% (19%)	0%
	Tiny-ImageNet	65	0.29	100% (17%)	0%
DistilBERT	Banking77	52	0.60	94.8% (23%)	0%
	CLINC150	61	0.48	97.3% (16%)	0%

Mean angle averages the principal angles between pairs of top-10 eigenspaces. Variance in shared directions uses the top-20 directions of the average covariance of all classes except class c. Both metrics use OAS-stabilized covariances. The distance columns use unshrunk full-sample covariances to form both targets and the class reference. CIFAR-10 is included only for this covariance analysis because TACO uses LDA without SCT when $C \leq 20$.

Mean angles range from $52°$ to $65°$, well below the random value of about $80°$. The main covariance directions are therefore closer than random directions, although they are not identical across classes. This matches SCT's design, which mixes shared structure with each class's own estimate instead of replacing it. Variance in shared directions ranges from 0.29 to 0.60, far above the random range of 0.02–0.04. The shared directions therefore carry a substantial part of each class's variance.

Before rotation, the SCT target is closer for every class in the six vision settings and reduces the distance by 13% to 22%. The corresponding rates are 94.8% and 97.3% in the two text settings, with distance reductions of 23% and 16%. After rotation, the closer rate falls to 0% in every setting. Because the eigenvalues are unchanged, the loss of this advantage in this rotation control shows that the smaller target distance depends on shared directions rather than eigenvalue magnitudes alone.

C.3 The SCT Target Remains Closer When Built from Few Samples

Table 4 gives the ideal version using all available samples. Table 5 repeats the comparison with limited construction data, the few-sample setting for which SCT is intended.

We use 20, 50, 100 or 200 randomly selected samples per eligible class to build the SCT cross-class target and the OAS scaled-identity target. All remaining samples form a disjoint reference set whose covariance approximates each class covariance for the distance comparison.

Table 5: **Target distance with few construction samples**. Each entry gives the percentage of eligible classes for which the SCT target is closer to the independent reference covariance than the OAS target. Parentheses give the mean distance reduction relative to OAS. Results are averaged over three random splits.

		Samples per class used to build the targets			
Encoder	**Dataset**	**20**	**50**	**100**	**200**
CLIP ViT-B/32	CIFAR-10	100% (14%)	100% (17%)	100% (19%)	100% (19%)
	CIFAR-100	100% (14%)	100% (19%)	100% (20%)	100% (20%)
	Tiny-ImageNet	100% (8%)	100% (10%)	100% (12%)	100% (12%)
ResNet-18	CIFAR-10	100% (10%)	100% (16%)	100% (19%)	100% (21%)
	CIFAR-100	100% (10%)	100% (15%)	100% (17%)	100% (17%)
	Tiny-ImageNet	100% (8%)	100% (12%)	100% (14%)	100% (15%)
DistilBERT	Banking77	100% (23%)	98.5% (24%)	N/A	N/A
	CLINC150	100% (15%)	99.8% (14%)	N/A	N/A

Within the construction set, the SCT target averages the other-class OAS covariances as in TACO, and the OAS target is a scaled identity with the same trace as the current class's covariance estimate. Distances are Frobenius distances to the unshrunk reference covariance. Reference sets contain at least 200 samples per vision class and 50 per text class. All vision and CLINC150 classes qualify. Banking77 uses 73 classes at 20 samples and 66 at 50. N/A means insufficient reference samples.

With only 20 construction samples per class, the SCT target is closer for every eligible class in all eight settings. Across all construction sizes, it is closer for every eligible class in the vision settings and for at least 98.5% of eligible classes in the text settings. It reduces the mean distance by 8% to 24%. The reduction tends to grow with construction size in the vision settings and remains similar at the two text sizes. The cross-class target therefore remains closer throughout the evaluated few-sample range.

C.4 How Per-Class Sample Size Affects Covariance Stability and Accuracy

Table 6 examines how the number of samples per class affects covariance stability and TACO accuracy with frozen CLIP ViT-B/32 features. Let N denote the number of sampled training examples per class. We split them into two disjoint groups of $\lfloor N/2 \rfloor$ examples and compute an unshrunk sample covariance from each group. A smaller Frobenius gap between these estimates means less sensitivity to sample choice.

Sampling variation in a sample covariance is expected to decrease approximately as $1/\sqrt{N}$. We therefore use $\sqrt{25/N}$ as the normalized inverse-square-root reference, which equals 0.50 at $N{=}100$ because quadrupling the sample count halves the expected variation. The accuracy rows use all N examples per class to construct TACO and report test accuracy without CTTA.

From $N{=}25$ to $N{=}400$, the expected relative gap reaches 0.25, while the measured gaps reach 0.24 on CIFAR-100 and 0.23 on Tiny-ImageNet. The measured decrease therefore closely follows the inverse-square-root trend. Over the same range, TACO accuracy rises from 59.45% to 75.34% on CIFAR-100 and from 55.85% to 69.28% on Tiny-ImageNet. These results quantify the covariance-estimation difficulty in the limited-data regime targeted by SCT.

Table 6: **Covariance stability and TACO accuracy across sample sizes**. Relative covariance gaps compare the estimates from the two data halves and are normalized to 1.00 at N=25. Lower gaps indicate greater stability.

	Samples per class N				
	25	**50**	**100**	**200**	**400**
Relative covariance gap ↓					
Expected from $1/\sqrt{N}$	1.00	0.71	0.50	0.35	0.25
CIFAR-100	1.00	0.68	0.47	0.33	0.24
Tiny-ImageNet	1.00	0.68	0.47	0.33	0.23
TACO accuracy (%) ↑					
CIFAR-100	59.45	63.18	69.84	73.71	75.34
Tiny-ImageNet	55.85	58.09	59.70	66.11	69.28

For each class, we compute the root-mean-square Frobenius gap over six random splits, then take the median across classes. Accuracy at each N comes from one run with N randomly selected examples per class.

C.5 SCT Outperforms Six Alternative Covariance Estimators

Table 7 compares SCT with six alternative covariance estimators under the same frozen features and QDA classifier.

All six alternatives use only the current class. Sample covariance applies no shrinkage. OAS (Chen et al., 2010) and Ledoit-Wolf (Ledoit & Wolf, 2004) shrink toward a scaled identity with different weights. Diagonal and constant-correlation targets (Schäfer & Strimmer, 2005) preserve feature variances and shrink correlations toward zero or their mean. The PPCA-style low-rank-plus-isotropic estimator (Tipping & Bishop, 1999) retains the leading directions and replaces the remaining eigenvalues by their mean. SCT instead uses the OAS covariances of the other classes as its shrinkage target.

Table 7: **QDA accuracy with different covariance estimators**. Accuracy (%) when the class covariance estimator changes. **Bold** and underline mark the best and second-best results. Shading marks SCT, and red arrows give its margin over the second-best result.

Encoder	**Dataset**	**Sample Cov.**	**OAS**	**LW**	**Diag.**	**Const. Corr.**	**Low-Rank +Iso.**	**SCT**
CLIP ViT-B/32	CIFAR-10	90.13	90.14	90.13	89.44	89.46	90.21	**90.40**↑0.19
	CIFAR-100	67.80	71.59	71.58	72.96	72.94	71.73	**75.38**↑2.42
	Tiny-ImageNet	58.98	63.67	63.76	66.89	66.89	66.80	**69.23**↑2.34
ResNet-18	CIFAR-10	84.99	85.49	85.48	84.23	84.21	83.87	**86.19**↑0.70
	CIFAR-100	55.43	65.50	65.51	66.20	66.09	65.10	**69.64**↑3.44
	Tiny-ImageNet	46.04	60.01	59.91	60.62	60.62	60.07	**66.16**↑5.54
DistilBERT	Banking77	79.35	80.68	79.71	80.00	79.97	76.88	**86.46**↑5.78
	CLINC150	87.80	86.18	85.87	86.76	86.78	83.16	**90.44**↑2.64

Sample Cov., OAS and LW denote sample covariance, Oracle Approximating Shrinkage and Ledoit-Wolf shrinkage. Diag., Const. Corr., Low-Rank+Iso. and SCT denote diagonal shrinkage, constant-correlation shrinkage, low-rank-plus-isotropic estimation and Structural Covariance Transfer. Diag., Const. Corr. and Low-Rank+Iso. report the highest test accuracy across the tested settings.

OAS and Ledoit-Wolf shrink toward the same scaled-identity target using different analytic weights. Their accuracies differ by at most 0.10% in the six vision settings and 0.97% in the two text settings, and neither matches SCT.

Structure estimated from the current class helps on CIFAR-100 and Tiny-ImageNet, but does not match cross-class transfer. Diagonal or constant-correlation shrinkage improves on OAS by up to 3.22% in these settings. On Banking77 and CLINC150, low-rank plus isotropic estimation falls

below sample covariance. This result is consistent with the difficulty of estimating leading directions in 768 dimensions from roughly one hundred examples per class.

SCT has the highest accuracy among the seven estimators in all eight settings. It leads the strongest alternative by 2.34% to 5.78% on CIFAR-100, Tiny-ImageNet, Banking77, and CLINC150. The margins narrow to 0.19% and 0.70% on CIFAR-10, where each class has 5000 samples. Across these datasets, the larger gains coincide with fewer samples relative to the feature dimension.

D Implementation Details

Default configuration. For datasets with $C \leq 20$ (CIFAR-10), we use LDA without fusion and set $\alpha_f = 0$. For $C > 20$ (CIFAR-100 and Tiny-ImageNet), we use QDA with dual-space fusion at $\alpha_f = 0.25$. SCT uses uniform cross-class averaging with $\gamma = 5000$. At test time, each sample has $V = 10$ views. CIFAR-10 uses resize-to-224 and horizontal flip. CIFAR-100 and Tiny-ImageNet use resize-to-232 and resize-to-236, respectively, followed by random crop to 224 and horizontal flip. The closed-form classifier is deterministic. Across five CTTA seeds, the standard deviation is at most 0.13%. For the main benchmarks, all hyperparameters are selected on CIFAR-100 and then held fixed.

Tuned configuration. TACO-Tune changes the client-side preprocessing, the covariance regularization, and the fusion weight, so the upload size is unchanged. Clients upsample inputs with bicubic interpolation and apply signed square-root normalization to the visual features before L2 normalization. Two settings depart from this. Tiny-ImageNet with ResNet-18 upsamples bilinearly, and CIFAR-10 with CLIP omits the square-root normalization. All six settings use QDA with OAS and SCT, with dual-space fusion at $\alpha_f = 0.15$. The SCT pseudo-count γ is 25000 on CIFAR-10 for both backbones, and $1500/3000$ on CIFAR-100 and $4000/6500$ on Tiny-ImageNet for ResNet-18/CLIP. Accuracies average three CTTA seeds with a standard deviation of at most 0.17%.

E Sensitivity Analysis

TACO has two tunable hyperparameters. The SCT pseudo-count γ controls how much covariance structure is shared across classes, and the fusion weight α_f balances the visual and text classifiers. We test whether each parameter can be fixed once and reused across datasets. Figure 4 shows both effects on the three main datasets.

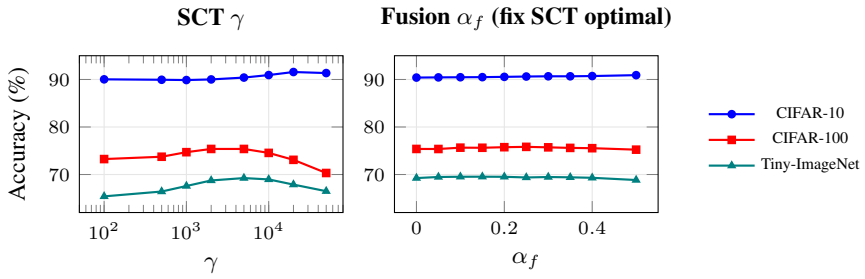

Figure 4: Hyperparameter sensitivity without CTTA. **Left:** SCT pseudo-count γ. **Right:** dual-space fusion weight α_f ($\alpha_f = 0$: visual-only, $\alpha_f = 1$: text-only).

γ controls the amount of cross-class transfer. A small γ leaves each estimate near its own OAS covariance. A larger γ moves it further toward the covariance shared by other classes. The left panel of Figure 4 shows this trade-off on the three image datasets. The shared default $\gamma = 5000$ attains the best or tied-best accuracy on CIFAR-100 and Tiny-ImageNet. CIFAR-10 is included only as a QDA diagnostic because TACO uses LDA without SCT when $C \leq 20$.

Fusion accuracy is insensitive to α_f. Figure 4 (right) shows a broad, nearly flat peak on CIFAR-100 and Tiny-ImageNet, where fusion is active. CIFAR-100 peaks at $\alpha_f = 0.25$, while Tiny-ImageNet peaks at $\alpha_f \in [0.10, 0.15]$. Every value in $[0.1, 0.3]$ stays within 0.10% of the best accuracy on both datasets. We therefore fix $\alpha_f = 0.25$ whenever fusion is active on the main benchmarks. TACO uses vi-

sual LDA without fusion on CIFAR-10, where a 10-dimensional text space adds little complementary information under the default representation.

The default γ=5000 transfers across encoders and modalities. Table 8 extends the comparison to eight combinations of encoder and dataset with CLIP ViT-B/32, ResNet-18, and DistilBERT. A value above OAS means that SCT improves on class-specific OAS shrinkage, while the column maximum shows the best value found for that setting.

Table 8: **SCT sensitivity across encoders and modalities**. Each column reports QDA accuracy (%) for ten values of the SCT pseudo-count γ and an OAS reference. The OAS row uses the class-specific OAS covariance estimates directly, with SCT disabled. All other classifier components are fixed. Shading marks the shared default γ=5000. **Bold** and underline mark the best and second-best SCT values in each column, including ties. CLIP denotes ViT-B/32, RN18 denotes ResNet-18, and BERT denotes DistilBERT. Text fusion and CTTA are disabled. The CIFAR-10 columns are QDA diagnostics because TACO uses LDA without SCT when $C \leq 20$.

	Vision, $C>20$				**CIFAR-10 diagnostics**		**Text**	
γ	**CLIP C100**	**CLIP Tiny**	**RN18 C100**	**RN18 Tiny**	**CLIP C10**	**RN18 C10**	**BERT Bank77**	**BERT CLINC150**
OAS	71.59	63.67	65.50	60.01	90.14	85.49	80.68	86.18
50	73.15	65.36	67.50	61.62	90.04	85.61	79.16	87.71
100	73.25	65.40	67.67	62.08	90.03	85.56	79.42	88.62
500	73.74	66.42	68.58	62.96	89.94	85.74	85.23	91.24
1K	74.70	67.58	69.46	64.25	89.88	85.76	87.11	**92.00**
2K	**75.38**	68.75	**70.24**	65.46	90.01	85.78	**87.86**	91.89
5K	**75.38**	**69.23**	69.64	**66.16**	90.40	86.19	86.46	90.44
10K	74.55	68.98	68.30	65.82	90.93	86.69	85.10	88.93
20K	73.08	67.89	66.45	64.93	**91.58**	**87.38**	81.72	87.69
50K	70.32	66.48	64.08	64.11	91.36	85.52	77.50	86.44
100K	69.00	65.46	62.99	63.44	83.62	65.56	76.07	85.96

At γ=5000, SCT exceeds OAS in all eight columns and is at most 1.56% below the best tested value in any column. The larger values 10000 and 20000 also exceed OAS in every column, although their gaps to the optimum widen on the text datasets. At smaller values, cross-class transfer can be too weak. At very large values, shared covariance can suppress useful class-specific structure. Both extremes reduce accuracy in several settings. These results support γ=5000 as a shared default across datasets, encoders, and modalities.

F Baseline Descriptions

Train-Then-Merge Methods. These methods train local ResNet-18 models on each client's data and aggregate them on the server in a single communication round.

- **O-FedAvg** (McMahan et al., 2017). The canonical one-shot baseline, each client runs local SGD on its shard and the server averages the resulting weight vectors once, with no further communication.
- **MA-Echo** (Su et al., 2023) moves local models toward common harmonized optima and then aggregates their parameters without server-side training.
- **DENSE** (Zhang et al., 2022) adopts data-free knowledge distillation in which the server distills an ensemble of local models using a generator that synthesizes pseudo samples.
- **Co-Boosting** (Dai et al., 2024) iteratively improves both the generated data and the ensemble model quality so that distillation and generation reinforce each other.
- **IntactOFL** (Zeng et al., 2024) treats every local model as an independent expert and combines them with a Mixture-of-Experts network whose routing is tuned by self-supervised

samples that approximate the local data distributions, so that all local knowledge is preserved intact.

- **FuseFL** (Tang et al., 2024) adopts a causal view of one-shot federated learning, decomposes each network into blocks, and progressively trains and fuses the blocks bottom-up to augment intermediate features without additional communication.
- **FedBEns** (Talpini et al., 2025) models local posteriors with mixtures of Laplace approximations and aggregates them at the server to construct a global Bayesian ensemble.
- **FAFI** (Zeng et al., 2025) addresses model inconsistency by training local models with unsupervised feature alignment and category-wise prototype learning, and then performing informativeness-aware feature fusion and prototype aggregation at the server.

Frozen-Encoder Methods. These methods use frozen pre-trained encoders. FALCON uploads synthesized feature sequences, and the others upload statistics. All methods marked with † in Table 1 are partition-invariant by construction.

- **FALCON** (Liu et al., 2026) trains local classifiers and autoregressive generators, then uploads the classifiers and synthesized feature sequences for server-side distillation. We run its official code with the two frozen encoders and report the mean over three partition seeds.
- **FedCGS** (Guan et al., 2025) aggregates class means and the global feature covariance from frozen pre-trained features to construct a Gaussian classifier without gradient-based optimization.
- **Fed3R** (Fanì et al., 2024) uploads the feature second moment $\mathbf{X}^\top\mathbf{X}$ and feature-label cross-moment $\mathbf{X}^\top\mathbf{Y}$. The server sums these matrices and solves a regularized least-squares classifier in closed form.
- **AFL** (He et al., 2025) computes analytic client matrices from frozen features and combines them with an absolute aggregation rule. The resulting classifier requires no gradient-based client optimization and is invariant to the client partition.
- **DeepAFL** (Tang et al., 2026) stacks analytic residual transformations on frozen features and learns them layer by layer using closed-form least-squares solutions. Its training protocol requires multiple rounds of client-server communication.
- **SAFLe** (Bacellar et al., 2026) quantizes each frozen feature into overlapping binary buckets, groups the bits at random, and sums one embedding lookup per group into the class logits, which reduces to a single-round least-squares problem under the AFL aggregation rule. Because no code is released, we reimplement it with its own input pipeline and a fixed ridge term.
- **GH-NB** (Turazza et al., 2026) fits a Gaussian-Head Naive Bayes classifier that assumes a diagonal class-conditional covariance.
- **GH-LDA** (Turazza et al., 2026) fits a Gaussian-Head Linear Discriminant Analysis classifier with a single pooled covariance shared across all classes.
- **GH-QDA** (Turazza et al., 2026) fits a Gaussian-Head Quadratic Discriminant Analysis classifier with a separate full covariance per class, enabling class-specific decision boundaries.
- **GH-Proto-Hyper** (Turazza et al., 2026) adds a lightweight low-rank residual head on top of the Gaussian-Head logits and refines it via knowledge distillation on samples synthesized in an estimated Fisher subspace. The refinement requires server-side optimization, taking roughly 25 min on CIFAR-100 in our setting (see Appendix G).
- **FedCOF** (Goswami et al., 2025) estimates per-class covariances from client-level class means, requiring only first-order client statistics at 0.20 MiB per client. It is not partition-invariant because the covariance estimator depends on how samples are split across clients.

Table 9: **Communication, runtime, and accuracy**. Accuracy for statistics-based methods is evaluated at Dirichlet α=0.1 with CLIP ViT-B/32. Upload sizes use float32 values. Runtime excludes shared feature extraction and measures client optimization for FAFI and server-side classifier construction otherwise. *TACO+CTTA includes caching V=10 test views. Shading marks TACO variants. In the accuracy block, **bold** and underline mark the best and second-best results.

(a) Per-client upload size (MiB)

Method	Upload	C10	C100	Tiny
FAFI (ICML'25)	Model parameters	∼45.00	∼45.00	∼45.00
AFL (CVPR'25)	$\hat{\mathbf{W}}_k$, $\mathbf{R}_k$, $\mathbf{C}_k$	2.02	2.20	2.39
FedCOF (NeurIPS'25)	$\boldsymbol{\mu}_c$	0.02	0.20	0.40
FALCON (AAAI'26)	Generator and classifier	78.25	78.61	78.99
GH-LDA (ICLR'26)	$\{\mathbf{N}_c, \mathbf{A}_c, \boldsymbol{\Sigma}^{\text{pool}}\}^{\text{vis}}$	0.52	0.70	0.89
GH-QDA (ICLR'26)	$\{\mathbf{N}_c, \mathbf{A}_c, \mathbf{S}_c\}^{\text{vis}}$	5.03	50.29	100.59
GH-Proto-Hyper (ICLR'26)	$\{\mathbf{N}_c, \mathbf{A}_c, \mathbf{S}_c\}^{\text{vis}}$	5.03	50.29	100.59
DeepAFL (ICLR'26)	Layer-wise statistics	—	139.41	162.47
SAFLe (CVPR'26)	Correlation blocks and cross-moments	—	∼114	∼119
TACO-Lite (text)	$\{\mathbf{N}_c, \mathbf{A}_c, \mathbf{S}_c\}^{\text{txt}}$	0.003	1.97	15.49
TACO (visual+text)	$\{\mathbf{N}_c, \mathbf{A}_c, \mathbf{S}_c\}^{\text{vis+txt}}$	5.03	52.26	116.08
TACO (visual+text) + CTTA*	$\{\mathbf{N}_c, \mathbf{A}_c, \mathbf{S}_c\}^{\text{vis+txt}}$	5.03	52.26	116.08

(b) Compute and accuracy

	Compute			Accuracy (%)		
Method	C10	C100	Tiny	C10	C100	Tiny
FAFI (ICML'25)	∼13.8 h	∼15.7 h	∼113.8 h	77.83	45.48	43.62
AFL (CVPR'25)	0.10 s	0.61 s	1.23 s	91.91	70.71	64.80
FedCOF (NeurIPS'25)	0.10 s	0.62 s	0.99 s	87.34	67.95	64.09
FALCON (AAAI'26)	—	—	—	90.01	70.93	67.12
GH-LDA (ICLR'26)	0.12 s	0.41 s	0.79 s	91.10	70.11	64.23
GH-QDA (ICLR'26)	0.21 s	1.99 s	3.99 s	89.52	72.97	65.67
GH-Proto-Hyper (ICLR'26)	∼30 s	∼25 min	∼80 min	91.54	73.94	69.29
DeepAFL (ICLR'26)	—	—	—	91.99	75.68	69.63
SAFLe (CVPR'26)	—	—	—	**92.09**	74.48	68.40
TACO-Lite (text)	0.01 s	0.37 s	2.56 s	88.15	71.47	65.71
TACO (visual+text)	0.23 s	3.11 s	7.61 s	91.60	75.84	69.38
TACO (visual+text) + CTTA*	∼8 min	∼8 min	∼23 min	92.06	**76.69**	**71.00**

G Communication Cost and Runtime

Hardware and software environment. All results in Table 9 are measured on a single NVIDIA A16 GPU (16 GB) hosted on server running Ubuntu 20.04. The software stack is Python 3.8.10, PyTorch 2.3.0 with CUDA 12.1 and cuDNN 8.9.2, torchvision 0.18.0, NumPy 1.24.4, SciPy 1.10.1, and openclip 2.32.0 for the CLIP ViT-B/32 backbone.

Table 9 compares TACO's per-client upload size, runtime and accuracy against baselines on CIFAR-10, CIFAR-100 and Tiny-ImageNet. Accuracy for statistics-based methods is evaluated under the CLIP ViT-B/32 backbone and Dirichlet α=0.1. Three observations follow.

(i) Closed-form methods share one compute regime. The measured single-round closed-form methods in Table 9 finish classifier construction in under eight seconds on the largest dataset. TACO's SCT and dual-space fusion add 20 ms to 3.6 sec over GH-QDA across the three benchmarks. GH-Proto-Hyper requires up to ∼80 min of server-side residual-head distillation, while FAFI requires up to ∼113.8 h of client-side ResNet-18 optimization.

(ii) TACO's bandwidth is set by QDA, not by the text anchor. Linear-boundary methods (FedCOF, GH-LDA and AFL) transmit approximately 0.02–2.4 MiB, while per-class QDA methods transmit 5–116 MiB because they send one $D \times D$ scatter matrix per class. Within TACO's single upload, the text-space statistics add 0.06%, 3.9%, and 15.4% relative to GH-QDA on CIFAR-10, CIFAR-100,

and Tiny-ImageNet. The visual scatter matrices therefore dominate the upload. On CIFAR-10, TACO uses LDA at the server but retains the same sufficient-statistics upload format, giving the same 5.03 MiB size as GH-QDA.

TACO constructs its classifier after a single upload, while DeepAFL requires repeated exchanges of analytic statistics across layers. SAFLe uploads about 114 and 119 MiB per client on CIFAR-100 and Tiny-ImageNet, more than TACO's single upload of 52.26 and 116.08 MiB, at lower accuracy. FALCON can alternatively upload its generator and local classifier, which amounts to 78 to 79 MiB on all three datasets because the payload is set by the generator size.

(iii) TACO-Lite offers a lightweight option for small label spaces. Transmitting only the C-dimensional text-space statistics yields payloads of 0.003–15.49 MiB. On CIFAR-10, TACO-Lite is approximately 670× smaller than AFL (0.003 versus 2.02 MiB) with accuracy 3.76% lower than AFL. On CIFAR-100, it exceeds AFL in accuracy (71.47% versus 70.71%) and uses less bandwidth (1.97 versus 2.20 MiB). Its $\mathcal{O}(C^3)$ text-space payload reaches 15.49 MiB on Tiny-ImageNet, making this option most useful for smaller label spaces.

Server memory as class and feature spaces grow. Table 10 reports memory during classifier construction using synthetic float32 statistics at different class counts C and feature dimensions D, with 64 classes processed at a time. For TACO, the visual and text branches are measured separately and the larger peak is reported. TACO and GH-QDA have the same measured peak in all four settings. At C=1000 and D=1024, the larger branch's statistics occupy 3.91 GiB and classifier construction peaks at 4.68 GiB.

Table 10: **Server memory at larger class counts and feature dimensions**. GPU memory (GiB). Stats reports the larger branch's resident sufficient statistics. The method columns report peak memory during classifier construction. Shading marks TACO.

Dataset and backbone	C	D	Stats	GH-QDA (ICLR'26)	TACO (Ours)
CIFAR-100, ViT-B/32	100	512	0.10	0.30	0.30
Tiny-ImageNet, ViT-B/32	200	512	0.20	0.39	0.39
Tiny-ImageNet, ViT-H/14	200	1024	0.78	1.55	1.55
ImageNet-1K, ViT-H/14	1000	1024	3.91	4.68	4.68

H Cross-Domain Generalization

This section tests whether TACO and its variant QDA+SCT, which omits text fusion, remain competitive beyond the main benchmarks, across feature modalities and larger label spaces. Six NLP datasets use frozen DistilBERT sentence embeddings. Seven vision datasets use frozen CLIP ViT-B/32 image embeddings and span fine-grained recognition, general object recognition, texture recognition, and scene recognition, with up to 397 classes. In every experiment, the encoder remains fixed and each classifier is built from aggregated feature statistics.

NLP benchmarks. We tokenize each document to at most 256 tokens and extract the final [CLS] state from frozen DistilBERT (Sanh et al., 2019). The resulting 768-dimensional feature is L2-normalized before computing class statistics. We then apply QDA+SCT directly in this text feature space.

Vision benchmarks across domains and class counts. We test visual QDA+SCT and TACO on seven datasets with 37 to 397 classes. Images are resized and center-cropped to 224×224, then encoded and L2-normalized by frozen CLIP ViT-B/32. Caltech-256 uses a fixed 60-image-per-class split that includes clutter, while SUN397 follows official split 01 with 50 images per class. Fed3R and AFL report their best regularized result.

Tables 11 and 12 extend the evaluation across text and vision features and to label spaces with up to 397 classes. SAFLe achieves the highest accuracy on AG News and DBPedia-14 and DeepAFL on 20Newsgroups. On HWU64, Banking77, and CLINC150, QDA+SCT exceeds DeepAFL by 1.76%, 2.37%, and 1.18% and SAFLe by 0.23%, 5.41%, and 1.17%, with only one communication round.

Among the ten datasets with $C \geq 37$, QDA+SCT alone outperforms all compared baselines on seven and both Fed3R and AFL on nine.

TACO ranks first on all seven vision datasets, exceeding DeepAFL by 0.41% to 5.18%, SAFLe by 2.00% to 8.27%, and QDA+SCT by 0.14% to 4.63%. These results demonstrate the strong performance of QDA+SCT across modalities and the additional gains of TACO across visual domains and larger label spaces.

Table 11: **Generalization to NLP datasets**. Accuracy (%) with frozen DistilBERT features (D=768, K=5, α=0.1). C and Train/cls denote classes and samples per class. NCM is Nearest Class Mean. Shading marks QDA+SCT. **Bold** and underline indicate the best and second-best results. Red arrows give our margin over the runner-up when best.

Dataset	C	Train/cls	NCM	Fed3R (ICML'24)	AFL (CVPR'25)	DeepAFL (ICLR'26)	SAFLe (CVPR'26)	QDA+SCT (Ours)
AG News (Zhang et al., 2015)	4	30K	84.61	90.68	90.71	90.95	**91.66**	89.34
DBPedia-14 (Zhang et al., 2015)	14	40K	92.00	98.91	98.98	99.02	**99.11**	98.47
20Newsgroups (Lang, 1995)	20	566	53.86	64.09	64.37	**64.39**	63.03	64.33
HWU64 (Liu et al., 2021)	68	303	62.42	85.71	85.24	85.58	87.11	**87.34**↑0.23
Banking77 (Casanueva et al., 2020)	77	130	47.56	84.90	82.69	84.06	81.02	**86.43**↑1.53
CLINC150 (Larson et al., 2019)	150	100	66.51	89.11	89.18	89.29	89.30	**90.47**↑1.17

Table 12: **Generalization across vision datasets and class counts**. Test accuracy (%). C and Train/cls denote classes and training images per class. Shading marks TACO. **Bold** and underline indicate the best and second-best results. Red arrows show the TACO margin over the second-best method when it ranks first.

Dataset	C	Train/cls	NCM	Fed3R (ICML'24)	AFL (CVPR'25)	DeepAFL (ICLR'26)	SAFLe (CVPR'26)	QDA+SCT	TACO (Ours)
Pets (Parkhi et al., 2012)	37	99	79.64	88.72	89.13	86.59	86.22	91.11	**91.77**↑0.66
DTD (Cimpoi et al., 2014)	47	80	64.52	73.67	73.88	75.96	73.14	72.07	**76.70**↑0.74
Aircraft (Maji et al., 2013)	100	67	37.74	48.63	48.78	50.32	44.12	49.32	**52.39**↑2.07
Food-101 (Bossard et al., 2014)	101	750	81.12	84.89	83.56	86.79	83.79	86.95	**87.79**↑0.84
Flowers-102 (Nilsback & Zisserman, 2008)	102	20	92.26	94.49	94.54	94.96	91.47	95.23	**95.37**↑0.14
Caltech-256 (Griffin et al., 2007)	257	60	87.01	89.09	87.61	90.10	88.68	90.02	**90.68**↑0.58
SUN397 (Xiao et al., 2010)	397	50	70.45	73.63	70.72	74.90	72.26	75.08	**75.75**↑0.67

I Same-Architecture Encoder Heterogeneity

TACO assumes that all clients use the same frozen encoder, so their statistics lie in one feature space. Different pre-trained weights can change what each feature coordinate represents, even when the encoder architecture and output dimension remain the same. We therefore evaluate QDA+SCT as an increasing number of clients use different pre-trained ResNet-18 weights.

We use one fixed Dirichlet split of CIFAR-100 across five clients with α=0.1. In S_0, every client uses standard ImageNet-1K weights (TorchVision maintainers and contributors, 2016). From S_1 to S_4, the first one to four clients use either weights from a lighter ImageNet-1K training recipe (Wightman et al., 2021) or Places365 weights (Zhou et al., 2018), while the test encoder remains standard ImageNet-1K. The mismatched share in Table 13 is the fraction of training samples held by these clients.

All methods lose accuracy as the mismatched share grows. QDA+SCT ranks first in every column by 2.05% to 5.45%. At S_4, it reaches 24.85% with the lighter ImageNet-1K weights and 24.88% with Places365 weights. On this fixed ResNet-18 partition, QDA+SCT maintains its advantage over the baselines, although its absolute accuracy decreases substantially under encoder-weight mismatch. Other architectures or feature dimensions need alignment before their statistics can be aggregated.

Table 13: **Accuracy under client encoder weight mismatch**. CIFAR-100 test accuracy (%). In S_k, the first k of five clients use the alternative weights named in each block. Shading marks QDA+SCT. **Bold** and underline indicate the best and second-best results. Red arrows show the QDA+SCT margin over the second-best method.

Method	S_0	S_1	S_2	S_3	S_4
Mismatched share	0.0%	20.8%	44.2%	63.8%	81.5%
Lighter ImageNet-1K weights					
NCM	54.12	40.54	30.09	19.06	9.59
AFL (CVPR'25)	58.56	53.39	41.42	33.65	22.80
QDA+OAS	65.50	46.66	32.07	19.71	9.90
GH-QDA (ICLR'26)	66.52	57.05	43.38	32.26	19.89
QDA+SCT	**70.47**↑3.95	**61.49**↑4.44	**48.83**↑5.45	**38.16**↑4.51	**24.85**↑2.05
Places365 weights					
NCM	54.12	42.06	31.33	21.14	11.01
AFL (CVPR'25)	58.56	53.15	41.18	33.47	22.56
QDA+OAS	65.50	49.35	35.13	23.36	13.93
GH-QDA (ICLR'26)	66.52	56.82	42.81	32.10	19.85
QDA+SCT	**70.47**↑3.95	**60.69**↑3.87	**48.14**↑5.33	**37.73**↑4.26	**24.88**↑2.32

J Effect of Encoder Choice and Pretraining

We study the shared frozen encoder along two axes. We first compare CLIP-family backbones. We then fix the architecture as ResNet-18 and vary its pretraining. Together, these experiments measure how TACO behaves across encoder capacity, pretraining source, and data domain.

J.1 Across CLIP-Family Backbones

We compare CLIP ViT-B/32, CLIP ViT-B/16, CLIP ViT-L/14, and OpenCLIP ViT-H/14 on CIFAR-10 and CIFAR-100 in Table 14.

Table 14: **Accuracy across CLIP-family backbones**. Test accuracy (%) on CIFAR-10 and CIFAR-100. Parameter counts are shown in parentheses. Shading marks TACO. **Bold** and underline indicate the best and second-best results. Red arrows show the TACO margin over the second-best method when it ranks first.

Backbone	**Dataset**	**FedCOF** (NeurIPS'25)	**GH-QDA** (ICLR'26)	**GH-Proto-Hyper** (ICLR'26)	**DeepAFL** (ICLR'26)	**SAFLe** (CVPR'26)	**TACO** (Ours)
CLIP ViT-B/32	CIFAR-10	87.34	89.52	91.54	91.99	**92.09**	92.06
(88M)	CIFAR-100	67.95	72.97	73.94	75.68	74.48	**76.69**↑1.01
CLIP ViT-B/16	CIFAR-10	91.44	92.84	**93.39**	93.21	93.25	93.33
(86M)	CIFAR-100	73.98	75.78	78.05	77.84	76.11	**79.46**↑1.41
CLIP ViT-L/14	CIFAR-10	95.70	96.18	96.70	96.42	96.20	**96.72**↑0.02
(304M)	CIFAR-100	81.28	81.84	83.21	82.57	81.77	**85.27**↑2.06
OpenCLIP ViT-H/14	CIFAR-10	97.17	97.63	**98.19**	97.75	97.64	98.16
(632M)	CIFAR-100	86.52	86.91	87.81	87.92	87.39	**89.43**↑1.51

Accuracy increases for every method from ViT-B/32 to ViT-H/14. TACO ranks first in five of the eight backbone and dataset combinations. On CIFAR-100, it exceeds the strongest baseline by 1.01% to 2.06% across the four backbones. On CIFAR-10, TACO ranks first with ViT-L/14, trails SAFLe by 0.03% with ViT-B/32, and trails GH-Proto-Hyper by 0.06% and 0.03% with ViT-B/16 and ViT-H/14. TACO remains competitive across the evaluated CLIP-family encoders and retains its largest gains on CIFAR-100.

J.2 Changing Pretraining with a Fixed ResNet-18

We next fix ResNet-18 and compare six weight settings for the shared encoder. They cover random initialization, Places365 scene pretraining (Zhou et al., 2018), three ImageNet-1K recipes with

increasing training budgets, and weak supervision on roughly one billion web images followed by ImageNet-1K fine-tuning (Yalniz et al., 2019). The light ImageNet-1K recipe (Wightman et al., 2021) uses low resolution and fewer epochs, the standard recipe (TorchVision maintainers and contributors, 2016) uses standard resolution for 90 epochs, and the strong recipe (Wightman et al., 2021) keeps this resolution but extends training to about 600 epochs with stronger augmentation. All clients use the same weights in each setting. Table 15 compares TACO with single-round GH-QDA (Turazza et al., 2026) and SAFLe (Bacellar et al., 2026) and multi-round DeepAFL (Tang et al., 2026).

Table 15: **Accuracy with a fixed ResNet-18 under different pretraining**. Test accuracy (%) on CIFAR-100 and Tiny-ImageNet. Shading marks TACO. **Bold** and underline indicate the best and second-best results. Red arrows show the TACO margin over the second-best method when it ranks first.

Pretraining	CIFAR-100				Tiny-ImageNet			
	GH-QDA (ICLR'26)	DeepAFL (ICLR'26)	SAFLe (CVPR'26)	TACO (Ours)	GH-QDA (ICLR'26)	DeepAFL (ICLR'26)	SAFLe (CVPR'26)	TACO (Ours)
Random init.	**26.99**	25.18	23.13	25.75	11.82	11.72	8.50	**13.95**↑2.13
Places365	51.40	**53.90**	53.17	53.10	42.91	**45.95**	45.81	45.24
IN1K, light	57.30	57.71	57.74	**60.23**↑2.49	54.99	57.16	56.56	**59.52**↑2.36
IN1K, standard	67.74	67.02	67.34	**70.21**↑2.47	61.83	62.23	61.93	**65.84**↑3.61
IN1K, strong	63.58	66.37	65.08	**67.17**↑0.80	60.07	63.59	61.94	**64.41**↑0.82
Web + IN1K	64.85	67.74	66.29	**68.96**↑1.22	59.34	63.56	62.43	**64.82**↑1.26

Pretraining raises the accuracy of all four methods far above random initialization. Among the five pre-trained settings, TACO exceeds GH-QDA on both datasets. The gains range from 1.70% to 4.11% on CIFAR-100 and from 2.33% to 5.48% on Tiny-ImageNet. TACO also exceeds DeepAFL and SAFLe in ten of the twelve dataset and pretraining combinations, with the two Places365 cases as exceptions. With random weights, TACO is 1.24% below GH-QDA on CIFAR-100 and 2.13% above it on Tiny-ImageNet. The standard ImageNet-1K recipe gives TACO its highest accuracy on both datasets.

K Robustness to Label Noise and Long-Tailed Class Counts

We test whether TACO remains ahead of the baselines under label noise and long-tailed class counts. On CIFAR-100, we replace 10% or 20% of labels with random incorrect labels, or resample class counts to 1:10 and 1:100 exponential profiles. Table 16 reports the resulting accuracy.

Table 16: **Accuracy under label noise and class imbalance**. Results (%) on CIFAR-100 with frozen CLIP ViT-B/32 and α=0.1. Min./cls is the smallest training-class size. Shading marks TACO. **Bold** and underline mark the best and second-best results. Red arrows give TACO's margin over the runner-up.

Condition	Min./cls	NCM	AFL (CVPR'25)	GH-QDA (ICLR'26)	DeepAFL (ICLR'26)	SAFLe (CVPR'26)	TACO (Ours)
Clean	500	61.31	70.71	72.97	75.68	74.48	**76.69**↑1.01
Label noise 10%	477	61.20	70.75	71.24	75.49	73.88	**76.15**↑0.66
Label noise 20%	476	60.83	70.71	70.74	75.08	73.56	**75.48**↑0.40
Long-tailed 1:10	50	60.16	57.87	59.68	67.27	63.09	**69.25**↑1.98
Long-tailed 1:100	5	56.39	43.60	46.99	46.82	45.39	**64.43**↑8.04

TACO ranks first in all five conditions. Its margin over the runner-up ranges from 0.40% to 1.01% on clean and noisy data, then reaches 1.98% and 8.04% at long-tail ratios of 1:10 and 1:100. Its margin over SAFLe is 1.92% to 2.27% on clean and noisy data and grows to 6.16% and 19.04% at the two ratios. At 1:100, the smallest class has five samples and TACO reaches 64.43% accuracy.

L CLIENT-SIDE TRAINING BUDGETS

A single upload does not by itself remove client-side training. Table 17 tests how far FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), and MOON (Li et al., 2021) progress when each client trains for 1, 10, or 50 epochs before its only upload. FAFI (Zeng et al., 2025) is one-shot but also trains client models under these budgets. AFL (He et al., 2025) uses a fixed one-epoch analytic client stage. The Gaussian-Head methods (Turazza et al., 2026) and TACO construct their classifiers directly from sufficient statistics. These fixed procedures do not vary with the iterative epoch budget.

Table 17: **Accuracy after one upload under different client-side training budgets**. Test accuracy (%) with five clients, ResNet-18, and Dirichlet α=0.1. Closed-form rows use their fixed procedures and span the three epoch columns. Shading marks TACO. **Bold** and underline indicate the best and second-best results. Red arrows show TACO's margin over the runner-up.

	CIFAR-10			CIFAR-100		
Method	**1 epoch**	**10 epochs**	**50 epochs**	**1 epoch**	**10 epochs**	**50 epochs**
Iterative methods restricted to one upload						
FedAvg (AISTATS'17)	34.75	65.31	71.39	2.59	19.75	38.83
FedProx (MLSys'20)	34.86	65.17	72.89	2.31	19.70	39.15
MOON (CVPR'21)	40.45	64.80	71.93	2.56	30.67	38.99
One-shot method with client training						
FAFI (ICML'25)	6.49	20.17	25.83	4.96	21.04	35.96
Closed-form one-shot methods						
AFL (CVPR'25)		80.75			58.56	
GH-NB (ICLR'26)		78.84			55.51	
GH-LDA (ICLR'26)		<u>86.05</u>			63.92	
GH-QDA (ICLR'26)		84.40			<u>66.52</u>	
GH-Proto-Hyper (ICLR'26)		85.74			64.05	
TACO (Ours)		**86.48**↑0.43			**70.26**↑3.74	

FAFI rises from 6.49% and 4.96% after one epoch to 25.83% and 35.96% after 50 epochs. After 50 epochs, FedProx reaches 72.89% on CIFAR-10 and 39.15% on CIFAR-100. TACO reaches 86.48% and 70.26% without local optimization, exceeding FedProx by 13.59% and 31.11%. It also exceeds GH-LDA by 0.43% on CIFAR-10 and GH-QDA by 3.74% on CIFAR-100. TACO therefore combines one upload with higher accuracy and no client training in these settings.

M DISCUSSION

TACO targets settings where data remain local, repeated communication is costly, and clients have limited resources for local training. Potential applications include cross-institutional image classification, enterprise intent classification, and compute-constrained edge learning. Appendix H evaluates task-oriented dialogue benchmarks, Appendix G reports the communication and accuracy trade-off, and Appendix L measures how client-side training budgets affect accuracy.

Clients transmit per-class counts, sums, and scatter matrices rather than raw data, individual features, model parameters, or gradients. Without shared gradients, TACO avoids gradient-sharing attacks such as DLG (Zhu et al., 2019). The aggregate moments may still reveal information when a client has very few samples in a class. Minimum-count thresholds, differential privacy, and secure aggregation are possible safeguards, but we do not evaluate them here.

N LIMITATIONS

TACO requires clients to use the same frozen feature coordinates. Statistics from different architectures or feature dimensions cannot be aggregated directly, and Appendix I shows that mismatched encoder weights reduce absolute accuracy. Communication is a second limitation. The payload scales as $\mathcal{O}(CD^2 + C^3)$ and reaches approximately 116 MiB per client on Tiny-ImageNet with CLIP ViT-B/32. Low-rank or sketched covariance representations could reduce this cost. Finally, transmitting aggregate moments does not provide a formal privacy guarantee.