PROGRESSIVE GEOMETRY ALIGNMENT FOR HETEROGENEOUS FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Statistical heterogeneity remains a fundamental challenge in federated learning, often leading to inconsistent representations across clients, thereby hindering global generalization. We identify an important optimization pathway behind this inconsistency: under conventional joint local optimization, classifiers co-adapt with heterogeneous client data, creating client-specific discriminative geometries that introduce an additional source of representation-gradient deviation. This motivates explicitly controlling the influence of client-specific decision structures on representation learning. Based on this principle, we propose FedVGA, which coordinates representation learning under a shared classifier-induced geometry and allows locally adapted decision structures to influence future representations only after server-side consolidation. FedVGA progressively refines this shared geometry, represented as a spherical Voronoi configuration, through a global–local–global loop consisting of Align-and-Adapt and Geometry-Confidence Aggregation. We theoretically characterize geometry-induced representation-gradient deviation and local-to-global geometry error propagation. Experiments under both label and feature heterogeneity demonstrate consistent performance gains, progressive geometric alignment, and pronounced improvements on boundary-sensitive samples.

1 INTRODUCTION

Federated learning (FL) enables multiple decentralized clients to collaboratively train a global model without sharing raw data (McMahan et al., 2017; Kairouz et al., 2021). Under the canonical FedAvg paradigm (McMahan et al., 2017), clients perform local optimization and periodically communicate model updates for global aggregation, making FL particularly attractive to privacy-sensitive applications such as mobile intelligence (Hard et al., 2018; Yang et al., 2018), healthcare (Rieke et al., 2020), and finance (Liu et al., 2023).

Despite its practical promise, FL is fundamentally challenged by statistical heterogeneity across clients (Kairouz et al., 2021; Zhu et al., 2021). Such heterogeneity commonly manifests as label shift and feature shift, where clients differ in class proportions or class-conditional data distributions (Hsu et al., 2019; Li et al., 2021b; Ye et al., 2023). These discrepancies drive local models toward client-specific solutions, resulting in inconsistent feature representations and biased decision boundaries, thereby degrading global generalization.

Numerous approaches have been developed to mitigate statistical heterogeneity in FL. Early studies primarily regularize local updates or correct client drift (Li et al., 2020; Karimireddy et al., 2020), while recent representation-alignment methods use prototypes or semantic anchors to promote cross-client consistency (Tan et al., 2022; Xu et al., 2023; Zhang et al., 2024). The growing emphasis on representation alignment suggests that learning a coherent representation space across clients is central to mitigating statistical heterogeneity. We further ask how such representation inconsistency arises during federated training.

Our key observation is that statistical heterogeneity is the underlying source of client divergence, but an important pathway through which it affects representation learning arises from the conventional joint local optimization process. Beyond producing predictive probabilities, a classifier plays a more fundamental role in representation learning: it induces a discriminative geometry in the representation space (Balestriero et al., 2019; Ma et al., 2023), which shapes representation gradients through

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.

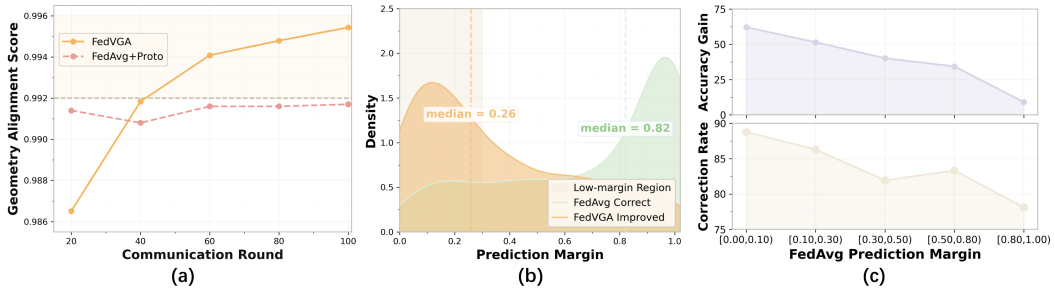

Figure 1: **Progressive geometric alignment and boundary-sensitive improvements of FedVGA.** **(a)** Evolution of cross-client geometry alignment during training. **(b)** Distributions of FedAvg prediction margins for samples correctly classified by FedAvg and samples misclassified by FedAvg but corrected by FedVGA. **(c)** Accuracy gain and correction rate conditioned on prediction margins, with smaller margins indicating greater predictive ambiguity under FedAvg. All results are obtained on the DomainNet. Detailed definitions are provided in Appendix E.

the classification objective and thereby serves as geometric supervision for representation formation. Under conventional federated optimization, however, local classifiers co-adapt with heterogeneous client data, causing the induced discriminative geometry to become client-specific. As a result, local representations are updated under different geometric supervision, inducing representation-gradient deviation and driving feature spaces toward different discriminative organizations before server aggregation (see Section 4). The server must therefore reconcile models whose representations have already been shaped by heterogeneous local supervision. In centralized learning, by contrast, data may still originate from heterogeneous underlying distributions, but their representations are jointly optimized toward a unified global discriminative geometry. As a result, feature updates remain aligned under shared geometric supervision, rather than being driven by independently adapted local decision structures. Taken together, this reveals client-specific geometric supervision as an important optimization pathway through which statistical heterogeneity is translated into cross-client representation inconsistency.

Motivated by this analysis, we aim to restore globally coordinated geometric supervision in federated representation learning while retaining the ability to adapt to distributed local data. To this end, we decouple representation formation from local classifier adaptation: representations are learned under a shared global discriminative geometry, while locally adapted decision structures are allowed to influence subsequent representation learning only after server-side consolidation. This differs from conventional representation–classifier decoupling (Kang et al., 2019), which primarily serves classifier re-balancing; here, decoupling serves as a federated optimization mechanism that controls when client-specific decision structures can enter global representation learning (see Appendix B.1). The classifier-induced discriminative geometry discussed above admits a natural Voronoi characterization. Prior geometric analyses show that the decision regions induced by linear classifiers admit Voronoi-type interpretations (Aurenhammer, 1987; Balestriero et al., 2019; Ma et al., 2023), providing an explicit geometric view of how classifier parameters organize the representation space. We argue that such geometry constitutes a natural alignment target for heterogeneous FL: it captures discriminative relations naturally formed through task optimization, with its joint configuration determining the decision boundaries. Building on this view, we construct a spherical ***Voronoi geometry*** to coordinate federated representation learning. More importantly, this explicit geometry naturally integrates cross-client alignment and progressive geometry refinement into the optimization process, without requiring a separate geometric alignment regularizer.

In this paper, we propose **Fed**erated **V**oronoi **G**eometry **A**lignment (**FedVGA**) for progressive geometric alignment under statistical heterogeneity. FedVGA constructs the spherical Voronoi geometry described above as the global alignment target and introduces an ***Align-and-Adapt*** mechanism to decouple representation learning from local geometry adaptation. To further consolidate heterogeneous local decision structures, we develop ***Geometry-Confidence Aggregation (GCA)***, which aggregates locally adapted geometries according to their geometric reliability. Together, these components form a global–local–global refinement loop: the current global geometry first coordinates local representation learning, the aligned representation space then provides a common basis for adapting local geometries, and their global consolidation produces the refined geometry that super-

vises the next cycle. Repeating this reciprocal process progressively aligns cross-client representations while simultaneously refining the shared discriminative geometry.

The training dynamics in Fig. 1 provide empirical support for this progressive geometric view. As training proceeds, FedVGA continuously improves cross-client geometry alignment, consistent with the reciprocal refinement between shared representations and global decision geometry. Moreover, its improvements are particularly pronounced for low-margin samples near decision boundaries: samples corrected by FedVGA concentrate more heavily in boundary-sensitive regions, where both the accuracy gain and correction rate are substantially higher. These observations suggest that FedVGA not only promotes cross-client representation consistency, but also progressively improves the discriminative organization of the shared representation space. Our main contributions are summarized as follows:

1. We propose FedVGA, a progressive geometry alignment framework that reorganizes federated optimization around globally coordinated geometric supervision. FedVGA constructs a Voronoi geometry as the structured global alignment target and controls when and how locally adapted decision geometry can re-enter representation learning through a global–local–global refinement process. This design prevents client-specific decision structures from directly perturbing feature learning while progressively aligning cross-client representations under a shared discriminative geometry.

2. We characterize an optimization pathway through which statistical heterogeneity affects representation learning: local classifier adaptation changes the geometric supervision received by representations. By isolating the resulting gradient deviation at the same representation state, we provide an analytical motivation for controlling when locally adapted decision structures enter subsequent representation learning.

3. We theoretically characterize geometry-induced representation-gradient deviation and local-to-global geometry error propagation. Extensive experiments under both label and feature heterogeneity demonstrate consistent performance gains, progressive cross-client geometry alignment, and pronounced improvements on boundary-sensitive samples, with ablations further validating the importance of controlling local geometry adaptation.

2 Related Work

2.1 Statistical Heterogeneity in Federated Learning

A substantial body of work addresses statistical heterogeneity in FL by improving federated optimization and aggregation. FedProx (Li et al., 2020) and FedDyn (Acar et al., 2021) regularize local optimization to restrict excessive local deviation, while SCAFFOLD (Karimireddy et al., 2020) corrects client drift through control variates. FedNova (Wang et al., 2020) reduces objective inconsistency through normalized aggregation, whereas FedBN (Li et al., 2021b) retains local batch normalization statistics to handle feature distribution shifts. Other studies explicitly model distribution discrepancies or improve representation robustness via distribution-aware aggregation, contrastive learning, or feature augmentation (Ye et al., 2023; Li et al., 2021a; Zhou & Konukoglu, 2023).

More recent methods promote cross-client representation consistency through shared class-level references. FedProto (Tan et al., 2022) and FedPAC (Xu et al., 2023) exploit class-wise prototypes for knowledge transfer and feature alignment, while FedTGP (Zhang et al., 2024) and FedLSA (Fu et al., 2025) introduce learnable global semantic targets. FedGLCL (Yan et al., 2025) further leverages language-driven semantics to construct a shared representation space. While these approaches differ substantially in their alignment mechanisms, they primarily address cross-client inconsistency by constructing or imposing shared semantic references. FedVGA instead studies this problem from the perspective of classifier-induced geometric supervision during federated representation learning. We discuss its relation to conventional representation–classifier decoupling in Appendix B.1.

2.2 Classifier-Induced Geometry

Voronoi and power diagrams provide classical tools for characterizing geometric partitions of representation spaces (Aurenhammer, 1987). Building on these foundations, geometric analyses of neural classifiers show that learned decision regions admit structured Voronoi-type interpretations

(Balestriero et al., 2019; Ma et al., 2022; 2023). These results establish an explicit connection between classifier parameters and the discriminative geometry that organizes representation space.

Building on this geometric perspective, FedVGA further constructs a spherical Voronoi geometry and uses it as a structured global alignment target for federated representation learning. Voronoi geometry has also been explored in FL. FedUD (Lei et al., 2025) explores Voronoi-based alignment in personalized FL, whereas FedVGA aims to establish a unified global geometry for cross-client representation alignment. Appendix B.3 provides a detailed comparison with FedUD.

3 METHODOLOGY

3.1 PRELIMINARIES

Federated Learning (FL). Consider a federated system with K clients, where each client k holds a local dataset $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_k|}$ sampled from its local distribution $P_k(x, y)$. Here, $|\mathcal{D}_k|$ denotes the number of local samples at client k. The number of samples belonging to class c at client k is denoted as $|\mathcal{D}_{k,c}|$. The objective of FL is to learn a global model Φ by minimizing the weighted empirical risk across all clients: $\min_\Phi \mathcal{L}(\Phi) = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \mathcal{L}_k(\Phi)$, where $|\mathcal{D}| = \sum_{k=1}^{K} |\mathcal{D}_k|$.

Geometric Interpretation of Linear Classifiers. A linear classifier admits a geometric interpretation through Voronoi diagrams. Under a particular relation between its weights and biases, the classifier-induced decision regions are equivalent to a Voronoi partition of the feature space.

Definition 1 (Voronoi Diagram (Aurenhammer et al., 2013))**.** *Given a set of sites* $\{\mathbf{c}_1, \ldots, \mathbf{c}_C\} \subset \mathbb{R}^d$*, the Voronoi cell associated with site* $\mathbf{c}_c$ *is* $\mathcal{V}_c = \left\{\mathbf{z} \in \mathbb{R}^d \mid \|\mathbf{z} - \mathbf{c}_c\|_2 \leq \|\mathbf{z} - \mathbf{c}_j\|_2,\ \forall j \neq c\right\}$*. The collection* $\{\mathcal{V}_1, \ldots, \mathcal{V}_C\}$ *constitutes the Voronoi diagram induced by these sites.*

Theorem 1 (Voronoi Diagram Reduction (Ma et al., 2022))**.** *A linear classifier parameterized by* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]^\top$ *and* $\mathbf{b}$ *partitions the feature space* $\mathbb{R}^d$ *into a Voronoi diagram with centers* $\{\mathbf{c}_1, \ldots, \mathbf{c}_C\}$ *given by* $\mathbf{c}_c = \frac{1}{2}\mathbf{w}_c, \quad$ *if* $\quad \mathbf{b}_c = -\frac{1}{4}\|\mathbf{w}_c\|_2^2, \quad c = 1, \ldots, C.$

Theorem 1 establishes a direct connection between classifier parameters and the geometry of their induced decision regions. Building on this connection, we further extend the reduction to the unit hypersphere to characterize the angular decision geometry. This explicit geometric representation enables classifier-induced decision structure to be shared and coordinated across clients, providing the basis for globally consistent geometric supervision in FedVGA.

3.2 FEDVGA: FEDERATED VORONOI GEOMETRY ALIGNMENT

We now formalize the proposed **FedVGA** framework. FedVGA is built on a simple principle: representation learning should be performed under globally coordinated geometric supervision, while locally adapted decision structures should influence future representation learning only after global consolidation. To realize this principle, FedVGA maintains a shared spherical Voronoi geometry and progressively refines it through a global–local–global process. The complete training procedure is summarized in Algorithm 1 in Appendix C.

Voronoi Geometry Construction. Building on Theorem 1, we first instantiate the classifier-induced Voronoi structure on the unit hypersphere, such that class decisions depend solely on angular similarity.

Let the global model be $\Phi = \mathcal{C} \circ \mathcal{F}$, where $\mathcal{F}$ is the feature extractor and $\mathcal{C}$ is a bias-free linear classifier with weights $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]^\top$. For an input $(\mathbf{x}, y)$, its feature representation is projected onto the unit hypersphere $\mathbb{S}^{d-1}$, and its prediction is performed by

$$\hat{y} = \arg\max_{c \in \{1,\ldots,C\}} (\boldsymbol{\mu}_c)^\top \mathbf{z}, \quad \text{where} \quad \mathbf{z} = \frac{\mathcal{F}(\mathbf{x})}{\|\mathcal{F}(\mathbf{x})\|_2}, \quad \boldsymbol{\mu}_c = \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2}. \tag{1}$$

Since $\|\mathbf{z} - \boldsymbol{\mu}_c\|_2^2 = 2 - 2\boldsymbol{\mu}_c^\top \mathbf{z}$, Eq. 1 is equivalent to assigning $\mathbf{z}$ to its nearest classifier direction on the hypersphere. This directly yields the following spherical Voronoi reduction, which serves as the geometric foundation of FedVGA.

Theorem 2 (Spherical Voronoi Diagram Reduction)**.** *Consider a bias-free linear classifier parameterized by* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C]^\top$*, whose predictions follow Eq. 1. Then the classifier partitions the hyperspherical feature space* $\mathbb{S}^{d-1}$ *into a spherical Voronoi diagram with Voronoi sites* $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C\}$ *given by* $\boldsymbol{\mu}_c = \frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2}$*. Specifically, the decision region associated with class* c *is*

$$\mathcal{V}_c = \left\{ \mathbf{z} \in \mathbb{S}^{d-1} \mid \boldsymbol{\mu}_c^\top \mathbf{z} \geq \boldsymbol{\mu}_j^\top \mathbf{z},\ \forall j \neq c \right\}. \tag{2}$$

Accordingly, we denote the resulting classifier-induced Voronoi geometry as $\mathcal{G} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C\}$. Importantly, FedVGA treats $\mathcal{G}$ as a joint configuration rather than a collection of independent class representatives. Each $\boldsymbol{\mu}_c$ defines a class-specific discriminative direction, while their joint configuration directly captures the classifier-induced organization of the feature space and its associated decision structure. This distinguishes our Voronoi geometry from conventional feature-mean prototypes, whose relative positions primarily reflect empirical class distributions. This structured Voronoi geometry is consequently used as the global alignment target in FedVGA.

Align-and-Adapt (A&A). To ensure that representation updates are consistently guided by globally coordinated geometric supervision, we introduce *Align-and-Adapt (A&A)*, which separates representation formation from local geometry adaptation and organizes them into a closed-loop geometric refinement process. This separation allows the shared geometry to evolve without letting client-specific adaptations directly perturb representation learning.

Align. At communication cycle t, let $\mathcal{G}^t = \{\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_C^t\}$ denote the current global Voronoi geometry. FedVGA first fixes $\mathcal{G}^t$ across all participating clients and optimizes only their feature extractors. For client k, the alignment objective is

$$\theta_k^{t+1} = \arg\min_{\theta_k} \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x},y)\in\mathcal{D}_k} -\log \frac{\exp\left(s(\boldsymbol{\mu}_y^t)^\top \mathbf{z}_k\right)}{\sum_{c=1}^{C} \exp\left(s(\boldsymbol{\mu}_c^t)^\top \mathbf{z}_k\right)}, \qquad \mathbf{z}_k = \frac{\mathcal{F}_{\theta_k}(\mathbf{x})}{\|\mathcal{F}_{\theta_k}(\mathbf{x})\|_2}, \tag{3}$$

where $s > 0$ is a scaling factor (Wang et al., 2017; 2018). Crucially, the geometry remains fixed throughout local feature optimization. Hence, despite heterogeneous local data, all participating clients update their representations under the same class-wise directions and decision partition specified by $\mathcal{G}^t$. The resulting feature extractors are then aggregated to obtain the updated global representation model θ^{t+1}.

With $\mathcal{G}^t$ fixed, local classifier adaptation cannot alter the geometric supervision received by representation updates. This removes the corresponding source of representation-gradient deviation, as formalized in Section 4. Under this shared supervision, Eq. 3 encourages each representation to favor its target-class site over competing sites in angular similarity, thereby guiding local representation learning through a common global decision geometry.

Adapt. After Align, the server redistributes the aggregated feature extractor θ^{t+1}, providing a common representation space for local geometry adaptation. Clients then adapt their decision structures to local data while keeping this extractor fixed, so that geometry adaptation does not directly modify the learned representations. Specifically, each client updates only its classifier-induced geometry:

$$\mathbf{W}_k^{t+1} = \arg\min_{\mathbf{W}_k} \frac{1}{|\mathcal{D}_k|} \sum_{(\mathbf{x},y)\in\mathcal{D}_k} -\log \frac{\exp\left(s\boldsymbol{\mu}_{k,y}^\top \mathbf{z}_k^{t+1}\right)}{\sum_{c=1}^{C} \exp\left(s\boldsymbol{\mu}_{k,c}^\top \mathbf{z}_k^{t+1}\right)}, \qquad \boldsymbol{\mu}_{k,c} = \frac{\mathbf{w}_{k,c}}{\|\mathbf{w}_{k,c}\|_2}, \tag{4}$$

where $\mathbf{z}_k^{t+1}$ is produced by the fixed feature extractor $\mathcal{F}_{\theta^{t+1}}$. The adapted classifier directions form the local geometry $\mathcal{G}_k^{t+1} = \{\boldsymbol{\mu}_{k,1}^{t+1}, \ldots, \boldsymbol{\mu}_{k,C}^{t+1}\}$, which characterizes the client-specific decision structure supported by the updated representation space. These local geometries are subsequently consolidated by GCA to construct the global geometry $\mathcal{G}^{t+1}$ for the next cycle.

The complete interaction can therefore be summarized as

$$\mathcal{G}^t \xrightarrow{\text{Align}} \{\theta_k^{t+1}\}_{k\in\mathcal{S}^{t,\text{Align}}} \xrightarrow{\text{Aggregation}} \theta^{t+1} \xrightarrow{\text{Adapt}} \{\mathcal{G}_k^{t+1}\}_{k\in\mathcal{S}^{t,\text{Adapt}}} \xrightarrow{\text{GCA}} \mathcal{G}^{t+1}. \tag{5}$$

Align-and-Adapt establishes a reciprocal refinement between representation and geometry: the global geometry provides unified supervision for representation learning, while the updated representation space provides a coordinated basis for subsequent geometry estimation. Repeating this cycle progressively couples cross-client representation alignment with global geometry refinement, without allowing local geometry adaptation to directly perturb the feature-alignment process.

Geometry-Confidence Aggregation (GCA). Local geometries obtained through Adapt are fitted to heterogeneous client data and may differ in reliability. Since their consolidation provides the shared supervision for the next cycle, this motivates accounting for the reliability of local estimates when constructing the global geometry. Drawing on the local-to-global error-propagation principle underlying the *Simplex Lemma* (Ding & Xu, 2020), GCA combines class-wise sample size with a geometry-confidence signal derived from the directional coherence of local representations. The geometric background and analysis of the resulting aggregation are provided in Appendix F.1 and Section 4, respectively.

For class c at client k, let $n_{k,c} = |\mathcal{D}_{k,c}|$. Using the aligned feature extractor $\mathcal{F}_{\theta^{t+1}}$, we compute

$$\bar{\mathbf{z}}_{k,c} = \frac{1}{n_{k,c}} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{k,c}} \mathbf{z}_{k,i}, \qquad \mathbf{z}_{k,i} = \frac{\mathcal{F}_{\theta^{t+1}}(\mathbf{x}_i)}{\|\mathcal{F}_{\theta^{t+1}}(\mathbf{x}_i)\|_2}, \qquad \rho_{k,c} = \|\bar{\mathbf{z}}_{k,c}\|_2. \tag{6}$$

Because $\rho_{k,c}^2$ contains a self-similarity contribution of $1/n_{k,c}$, it can overestimate directional concentration for small classes; in particular, $\rho_{k,c} = 1$ whenever $n_{k,c} = 1$. We therefore define the geometry-coherence score

$$\kappa_{k,c} = \frac{n_{k,c}\rho_{k,c}^2 - 1}{n_{k,c} - 1} \quad (n_{k,c} > 1), \qquad \kappa_{k,c} = 0 \quad (n_{k,c} = 1). \tag{7}$$

For $n_{k,c} > 1$, $\kappa_{k,c}$ equals the average pairwise cosine similarity between distinct class representations, providing an observable measure of directional coherence. We use this to construct geometry confidence for aggregation. Its statistical interpretation is given in Section 4 and Appendix F.3.

GCA converts this coherence into the geometry-confidence factor $r_{k,c} = 1 + \lambda\kappa_{k,c}$, where $\lambda \in [0, 1)$ is a hyperparameter, and defines the aggregation weight as

$$\alpha_{k,c} = \frac{n_{k,c} r_{k,c}}{\sum_{j \in \mathcal{S}_c^{t,\text{Adapt}}} n_{j,c} r_{j,c}} = \frac{n_{k,c}(1 + \lambda\kappa_{k,c})}{\sum_{j \in \mathcal{S}_c^{t,\text{Adapt}}} n_{j,c}(1 + \lambda\kappa_{j,c})}. \tag{8}$$

Thus, $n_{k,c}$ provides class-wise statistical support, while $r_{k,c}$ modulates this support according to geometric confidence. Since $\kappa_{k,c} \in [-1, 1]$ and $\lambda < 1$, $r_{k,c} > 0$; when $\lambda = 0$, Eq. 8 exactly reduces to conventional class-wise sample-size aggregation. For $\mathcal{S}_c^{t,\text{Adapt}} \neq \emptyset$, the Voronoi geometries are aggregated and projected back onto the hypersphere:

$$\boldsymbol{\mu}_c^{t+1} = \frac{\widetilde{\boldsymbol{\mu}}_c^{t+1}}{\|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2}, \qquad \widetilde{\boldsymbol{\mu}}_c^{t+1} = \sum_{k \in \mathcal{S}_c^{t,Adapt}} \alpha_{k,c} \boldsymbol{\mu}_{k,c}^{t+1}. \tag{9}$$

If $\mathcal{S}_c^{t,\text{Adapt}} = \emptyset$ or $\|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2 \leq \varepsilon$, we retain the previous site $\boldsymbol{\mu}_c^t$, where $\varepsilon > 0$ is a small numerical threshold. The resulting Voronoi sites form $\mathcal{G}^{t+1} = \{\boldsymbol{\mu}_c^{t+1}\}_{c=1}^C$, which serves as the shared Voronoi geometry for the next *Align-and-Adapt* cycle.

4 Theoretical Analysis

We analyze geometry-induced representation-gradient deviation and local-to-global geometry error propagation in the global–local–global process. We also characterize how representation synchronization affects local classifier adaptation, connecting the two stages of A&A. Detailed proofs are provided in Appendix F.

Effect of Align-and-Adapt. Let $\mathcal{G}^t = \{\boldsymbol{\mu}_c^t\}_{c=1}^C$ denote the global Voronoi geometry at communication cycle t. For a normalized representation $\mathbf{z}$ with label y, the feature gradient induced by a geometry $\mathcal{G}^t$ is $\mathbf{g}(\mathbf{z}, y; \mathcal{G}^t) = \nabla_{\mathbf{z}} \ell(\mathbf{z}, y; \mathcal{G}^t) = s\left(\sum_{c=1}^C p_c(\mathcal{G}^t)\boldsymbol{\mu}_c^t - \boldsymbol{\mu}_y^t\right)$, where $p_c(\mathcal{G}^t)$ is the corresponding softmax probability.

Under joint feature–classifier optimization, let $\mathcal{G}_{k,e}^t$ denote the local classifier geometry of client k at local step e. At the same representation state $\mathbf{z}_{k,e}$, the feature gradient can be decomposed with respect to the shared geometry $\mathcal{G}^t$ as

$$\mathbf{g}_{k,e}^{\text{joint}} = \underbrace{\mathbf{g}(\mathbf{z}_{k,e}, y; \mathcal{G}^t)}_{\mathbf{g}_{k,e}^g} + \underbrace{\left[\mathbf{g}(\mathbf{z}_{k,e}, y; \mathcal{G}_{k,e}^t) - \mathbf{g}(\mathbf{z}_{k,e}, y; \mathcal{G}^t)\right]}_{\mathbf{e}_{k,e}}, \tag{10}$$

where $\mathbf{e}_{k,e}$ measures the change in geometric supervision induced by local classifier adaptation.

Theorem 3 (Geometry-Induced Representation-Gradient Deviation). *Let $\delta_{k,e} = \max_c \|\boldsymbol{\mu}_{k,c}^{t,e} - \boldsymbol{\mu}_c^t\|_2$ denote the local-to-global geometry discrepancy. For unit-normalized representations and Voronoi sites, the feature-gradient deviation under joint optimization satisfies*

$$\|\mathbf{e}_{k,e}\|_2 = \left\| \mathbf{g}_{k,e}^{\text{joint}} - \mathbf{g}_{k,e}^{g} \right\|_2 \leq s(2+s)\delta_{k,e}. \tag{11}$$

Theorem 3 shows that local classifier-geometry drift can induce representation-gradient deviation, whose upper bound scales linearly with $\delta_{k,e}$. In the *Align* stage, however, $\mathcal{G}_{k,e}^t = \mathcal{G}^t$ throughout local feature optimization, and therefore $\delta_{k,e} = 0 \Longrightarrow \mathbf{e}_{k,e} = 0$.

During *Adapt*, the aggregated feature extractor is redistributed and held fixed, providing a common representation basis for local geometry adaptation (Proposition 2, Appendix F.2). Crucially, locally adapted geometries influence subsequent representation learning only after being consolidated into the shared global geometry $\mathcal{G}^{t+1}$. Thus, A&A allows heterogeneous local data to refine the geometric supervision while ensuring that representation updates are guided by a shared, globally consolidated geometry.

Effect of Geometry-Confidence Aggregation. We analyze the directional basis of geometry confidence and characterize how confidence-aware weighting affects local-to-global geometry error propagation.

Proposition 1 (Directional Basis of Geometry Confidence). *Let $\{\mathbf{z}_{k,i}\}$ be i.i.d. unit-normalized representations drawn from class c at client k, and let $\mathbf{z}$ and $\mathbf{z}'$ denote two independent draws from the same distribution. Define the population directional agreement as $\tau_{k,c} = \mathbb{E}[\mathbf{z}^\top \mathbf{z}']$. For $n_{k,c} > 1$,*

$$\kappa_{k,c} = \frac{1}{n_{k,c}(n_{k,c}-1)} \sum_{i \neq j} \mathbf{z}_{k,i}^\top \mathbf{z}_{k,j}, \qquad \mathbb{E}[\kappa_{k,c}] = \tau_{k,c}. \tag{12}$$

Proposition 1 shows that $\kappa_{k,c}$ measures the directional agreement between distinct class representations and provides an unbiased estimate of its population counterpart. It therefore provides an observable geometric basis for the confidence factor $r_{k,c} = 1 + \lambda\kappa_{k,c}$. Importantly, $\kappa_{k,c}$ measures coherence among class representations rather than directly estimating the error of the learned Voronoi site.

We next analyze how GCA affects the propagation of local geometry errors into the global estimate.

Theorem 4 (Local-to-Global Geometry Error Propagation). *Let $\boldsymbol{\mu}_c^\star$ denote a reference class-c site and define $\epsilon_{k,c} = \|\boldsymbol{\mu}_{k,c}^{t+1} - \boldsymbol{\mu}_c^\star\|_2$. For the GCA weights $\{\alpha_{k,c}\}$,*

$$\left\|\widetilde{\boldsymbol{\mu}}_c^{t+1} - \boldsymbol{\mu}_c^\star\right\|_2 \leq \sum_{k \in \mathcal{S}_c^{t,Adapt}} \alpha_{k,c}\epsilon_{k,c}. \tag{13}$$

Theorem 4 characterizes how GCA weights govern the contribution of local site errors to the aggregated geometry. Appendix F.3 further accounts for spherical normalization and propagates these errors to representation-gradient deviation in the next *Align* stage (Corollary 1). It also gives a sufficient condition under which GCA yields an error-propagation bound no larger than class-wise sample-size aggregation.

5 Experiments

5.1 Experimental Setup

Datasets and Evaluation Protocol. We evaluate FedVGA under both label and feature heterogeneity. For label shift, CIFAR-10/100 (Krizhevsky, 2009) are partitioned across clients using a Dirichlet distribution $\mathrm{Dir}(\alpha)$ over class proportions, following common FL protocols (Oh et al., 2021; Zhang et al., 2022; Acar et al., 2021). We consider $K \in \{50, 100, 200\}$ clients, with 50% client participation in each communication round. For feature shift, we use Office-Caltech-10 (Gong et al., 2012) and DomainNet (Peng et al., 2019), where each domain is treated as one client, resulting in 4 and 6 clients, respectively. All methods are evaluated using a single global model on a common test set which contains samples from all clients.

Table 1: **Test accuracy (%)** under feature-shift settings on DomainNet: C(Clipart), I(Infograph), P(Painting), Q(Quickdraw), R(Real), S(Sketch), and Office-Caltech-10: A(Amazon), C(Caltech), D(DSLR), W(Webcam). Avg. denotes the global test accuracy over all domains.

Method	DomainNet							Office-Caltech-10				
	C	I	P	Q	R	S	Avg.	A	C	D	W	Avg.
FedAvg	57.23	31.48	44.27	66.50	61.02	51.31	53.46 ± 0.87	78.12	56.44	43.75	66.10	64.96 ± 0.27
FedNova	59.75	26.85	44.27	59.50	62.15	58.12	53.46 ± 0.15	52.08	58.67	43.75	52.54	54.53 ± 0.26
FedProx	59.75	22.22	43.23	70.50	65.54	54.45	54.82 ± 0.40	76.04	45.33	43.75	54.23	57.87 ± 0.37
MOON	62.89	25.93	43.75	58.00	60.45	61.78	53.85 ± 0.23	75.00	64.00	68.75	71.19	69.29 ± 0.08
FedNH	64.78	35.19	44.79	69.50	59.32	53.93	55.89 ± 0.18	79.69	65.33	71.88	72.88	72.05 ± 0.43
FedFA	60.38	27.78	42.71	65.50	68.36	57.59	55.50 ± 0.08	70.31	76.00	71.88	66.10	72.44 ± 0.67
FedLSA	46.54	23.15	38.54	62.00	62.15	48.17	48.59 ± 0.33	52.08	81.33	62.50	62.71	66.93 ± 0.12
FedETF	73.58	38.89	46.35	70.00	71.19	57.07	60.66 ± 0.28	75.00	74.67	62.50	66.10	73.03 ± 0.52
FedVGA (Ours)	69.81	35.19	55.21	77.50	80.23	65.97	$\mathbf{66.02 \pm 0.39}$	71.35	83.56	71.88	74.58	$\mathbf{77.16 \pm 0.23}$

Baselines. We compare FedVGA with representative global FL methods: **FedAvg** (McMahan et al., 2017), **FedProx** (Li et al., 2020), **FedNova** (Wang et al., 2020), **MOON** (Li et al., 2021a), **FedNH** (Dai et al., 2023), **FedFA** (Zhou & Konukoglu, 2023), **FedLSA** (Fu et al., 2025), and **FedETF** (Li et al., 2023b). For label heterogeneity, we additionally include **FedLC** (Zhang et al., 2022). Methods requiring client-specific or personalized parameters are excluded due to the different single-global-model evaluation protocol.

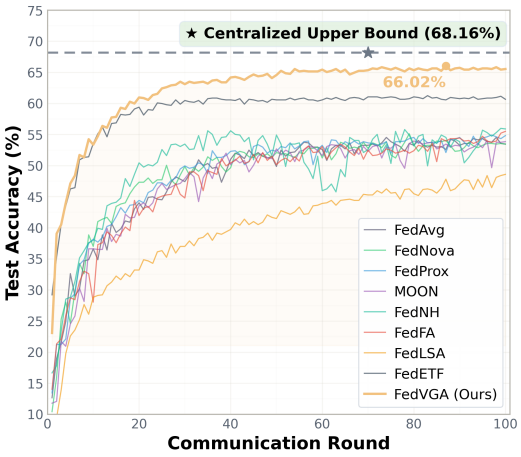

Figure 2: **Test accuracy (%) evolution** of different federated learning methods on DomainNet. The dashed line denotes the performance of centralized training as a reference.

Implementation Details. All methods use ResNet-18 (He et al., 2016) on CIFAR-10/100 and AlexNet (Krizhevsky et al., 2012) on Office-Caltech-10 and DomainNet, and are trained for 100 communication rounds. For FedVGA, one communication round denotes one complete *Align-and-Adapt* cycle. Both the *Align* and *Adapt* stages use the same or fewer local epochs as the local update of the competing methods. The scaling factor is set to $s = 40$, while the geometry-confidence coefficient $\lambda \in [0, 1)$ is specified separately for each dataset. Dataset-specific λ values and additional implementation details are provided in Appendix D.

5.2 Comparison with State-of-the-Art Methods

Tables 1 and 3 compare FedVGA with representative federated learning methods under feature and label heterogeneity. FedVGA achieves consistently competitive or superior performance across datasets and settings. For example, on CIFAR-100, FedVGA improves over FedAvg (McMahan et al., 2017) by **13.49** and **11.30** percentage points under $(K, \alpha) = (50, 0.1)$ and $(200, 0.3)$, respectively. Under feature shift, FedVGA achieves **66.02%** on DomainNet and **77.16%** on Office-Caltech-10, outperforming FedETF (Li et al., 2023b) by **5.36** and **4.13** points. As shown in Fig. 2, FedVGA maintains a clear advantage throughout training and reaches approximately **96.86%** of centralized performance on DomainNet, leaving only a **2.14**-point gap.

Table 2: **Ablation study** of the key components in FedVGA. For CIFAR-100, results are reported under $(K, \alpha) = (100, 0.1)$.

Variant	CIFAR-100	DomainNet	Office-Caltech-10
w/o VG	30.25	62.71	70.87
w/o A&A	27.59	55.32	61.61
w/o GCA	36.84	64.01	77.01
FedVGA	**37.93**	**66.02**	**77.16**

Table 3: **Test accuracy (%)** under label-shift settings on CIFAR-100 and CIFAR-10. Each setting is denoted by (K, α), where K is the number of clients and α is the Dirichlet concentration parameter controlling the degree of label heterogeneity; a smaller α indicates more severe label shift.

Method	CIFAR-100				CIFAR-10		
	(50,0.1)	(100,0.1)	(100,0.5)	(200,0.3)	(100,0.1)	(100,0.5)	(200,0.3)
FedAvg	27.45 ± 0.12	26.29 ± 0.08	25.77 ± 0.25	22.36 ± 0.28	45.30 ± 0.18	60.90 ± 0.64	55.37 ± 0.47
FedNova	28.49 ± 0.08	27.33 ± 0.15	27.68 ± 0.52	23.43 ± 0.18	46.79 ± 0.19	60.45 ± 0.37	56.30 ± 0.14
FedProx	26.95 ± 0.22	25.47 ± 0.09	26.89 ± 0.25	24.31 ± 0.74	50.11 ± 0.44	59.88 ± 0.12	55.24 ± 0.12
FedFA	30.02 ± 0.23	26.45 ± 0.20	28.64 ± 0.08	25.21 ± 0.14	49.27 ± 0.06	66.68 ± 0.33	60.32 ± 0.58
FedLC	29.88 ± 0.14	23.38 ± 0.27	27.42 ± 0.84	24.28 ± 0.44	55.86 ± 0.28	60.82 ± 0.28	55.69 ± 0.09
FedNH	33.46 ± 0.28	29.38 ± 0.54	30.82 ± 0.82	31.18 ± 0.26	57.00 ± 0.18	66.31 ± 0.18	63.24 ± 0.44
FedETF	35.56 ± 0.56	28.11 ± 0.33	37.67 ± 0.26	21.51 ± 0.42	55.19 ± 0.48	71.10 ± 0.29	56.81 ± 0.27
FedLSA	36.24 ± 0.17	30.00 ± 0.12	32.61 ± 0.31	22.23 ± 0.19	56.17 ± 0.23	67.42 ± 0.24	57.87 ± 0.12
FedVGA (Ours)	$\mathbf{40.94 \pm 0.24}$	$\mathbf{37.93 \pm 0.15}$	$\mathbf{42.88 \pm 0.46}$	$\mathbf{33.66 \pm 0.33}$	$\mathbf{60.61 \pm 0.67}$	$\mathbf{77.07 \pm 0.26}$	$\mathbf{70.13 \pm 0.15}$

When Does Geometry Alignment Help? We examine how FedVGA's advantage over FedAvg varies with label heterogeneity on CIFAR-100 (K=100) (Zhou & Konukoglu, 2023; Xu et al., 2023), including an IID reference alongside the Dirichlet label-skew settings. As shown in Fig. 3, FedVGA achieves larger gains over FedAvg in all evaluated label-skewed settings than under IID partitioning. Although the gains are not monotonic in the degree of label heterogeneity, the gain narrows to 7.77 percentage points at $\alpha = 3.0$, while remaining above the IID gain of 3.70 points. Viewed against the IID reference, this contrast suggests that FedVGA helps mitigate the accuracy loss associated with label heterogeneity, highlighting its value in heterogeneous federated learning settings.

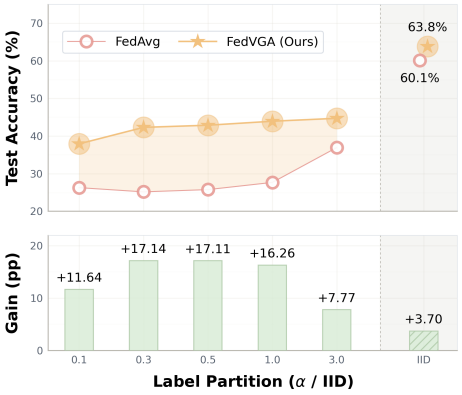

Figure 3: **Test accuracy (%)** (top) and gains over FedAvg in percentage points (bottom) on CIFAR-100 with $K = 100$ clients across Dirichlet label-skew settings and an IID reference.

5.3 Ablation Study

We evaluate three variants: (1) ***w/o VG***, which replaces the Voronoi geometry with conventional class-wise prototypes; (2) ***w/o A&A***, which retains the Voronoi geometry but jointly optimizes the feature extractor and classifier; and (3) ***w/o GCA***, which replaces geometry-confidence weighting with class-wise sample-size aggregation. As shown in Table 2, removing A&A produces the largest accuracy drop on all three benchmarks, supporting the role of separating representation learning from local geometry adaptation. Detailed global-local-global loop ablations in Appendix G further support the effectiveness of the complete Align-and-Adapt loop.

5.4 Geometry Alignment and Boundary-Sensitive Analysis

To further characterize the behavior of FedVGA beyond predictive accuracy, we analyze cross-client geometry alignment together with the distribution of samples affected by the method. As shown in Fig. 1(a), FedVGA exhibits a progressive increase in geometry alignment throughout training, indicating increasingly consistent representations at the decision-geometry level. From a complementary perspective, Fig. 1(b) shows that samples corrected by FedVGA are more concentrated in low-margin regions. The margin-conditioned results in Fig. 1(c) further show larger accuracy gains in low-margin regions while maintaining high correction rates across margin intervals. Fig. 1 provides complementary evidence of progressive decision-geometry alignment and improved predictions on samples that are ambiguous under the baseline model.

Together with the performance comparisons and ablations, these findings support FedVGA's design, in which locally adapted decision structures influence subsequent representation learning only after global consolidation. Further discussion and limitations are provided in Appendix A.

AI USE STATEMENT

We used generative AI tools to provide feedback on the formulation and scope of theoretical claims, experimental comparisons, and the interpretation of results. We also used these tools to assist with literature searches, manuscript organization, and the drafting and revision of text for clarity and consistency. We are responsible for verifying all AI-assisted content and for the accuracy and integrity of the final manuscript, including its theoretical claims, experimental results, and references.

REPRODUCIBILITY STATEMENT

The objectives and update rules of FedVGA are specified in Section 3, with the training procedure summarized in Algorithm 1. Section 4 and the additional theoretical analysis in the appendix state the assumptions and provide proofs for the analytical results. Section 5.1 and the experimental-details (Appendix D) describe the datasets, client partitions, model architectures, and training configurations. Appendix E defines the evaluation metrics and sample-grouping procedures used in Fig. 1. The implementation code is included in the supplementary materials.

REFERENCES

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

Franz Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Franz Aurenhammer, Rolf Klein, and Der-Tsai Lee. *Voronoi diagrams and Delaunay triangulations*. World Scientific Publishing Company, 2013.

Randall Balestriero, Romain Cosentino, Behnaam Aazhang, and Richard Baraniuk. The geometry of deep networks: Power diagram subdivision. *Advances in Neural Information Processing Systems*, 32, 2019.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021.

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

Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7314–7322, 2023. doi: 10.1609/aaai.v37i6.25891.

Hu Ding and Jinhui Xu. A unified framework for clustering constrained data without locality property. *Algorithmica*, 82(4):808–852, 2020.

Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16648–16656, 2025.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073. IEEE, 2012.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

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*, pp. 770–778, 2016.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

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*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Mingxi Lei, Meng Ding, Chunwei Ma, Ziyun Huang, Mingchen Gao, and Jinhui Xu. Unifying domain gap in federated learning: A geometric approach. *Neurocomputing*, 617:128813, 2025.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021a.

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

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

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning*, pp. 19767–19788. PMLR, 2023a.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5319–5329, 2023b.

Tao Liu, Zhi Wang, Hui He, Wei Shi, Liangliang Lin, Ran An, and Chenhao Li. Efficient and secure federated learning for financial applications. *Applied Sciences*, 13(10):5877, 2023.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.

Chunwei Ma, Ziyun Huang, Mingchen Gao, and Jinhui Xu. Few-shot learning as cluster-induced voronoi diagrams: A geometric approach. *arXiv preprint arXiv:2202.02471*, 2022.

Chunwei Ma, Zhanghexuan Ji, Ziyun Huang, Yan Shen, Mingchen Gao, and Jinhui Xu. Progressive voronoi diagram subdivision enables accurate data-free class-incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.

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

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, Sebastien Ourselin, Micah Sheller, Ronald M Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):119, 2020.

Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. *arXiv preprint arXiv:2204.13399*, 2022.

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

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5265–5274. IEEE, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33:7611–7623, 2020.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Salman Khan, Lei Zhu, and Yong Liu. On the importance of language-driven representation learning for heterogeneous federated learning. In *International Conference on Learning Representations*, 2025.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. In *International Conference on Machine Learning*, pp. 39879–39902. PMLR, 2023.

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

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International conference on machine learning*, pp. 26311–26329. PMLR, 2022.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

Tianfei Zhou and Ender Konukoglu. Fedfa: Federated feature augmentation. *arXiv preprint arXiv:2301.12995*, 2023.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

A Discussion and Limitations

FedVGA can be viewed as a federated decomposition of the coordination between representation learning and classifier adaptation in centralized learning. In centralized training, data from heterogeneous sources contribute to a common classifier, whose decision geometry in turn supervises representation learning. Under joint local optimization in FL, this interaction occurs separately on each client: locally adapted decision structures can shape representations before global consolidation. FedVGA takes centralized learning as an organizational reference, seeking to preserve the globally coordinated character of geometric supervision while allowing adaptation to distributed data.

From this perspective, decoupling separates the collection of local discriminative information from its use as shared supervision. Local adaptation contributes information about the decision structure supported by each client's data; global consolidation determines the common geometry through which this information influences subsequent representation learning. Geometry therefore serves as an interface between distributed adaptation and globally coordinated supervision. This interpretation motivates a broader question for federated optimization: which learned structures should be globally consolidated before they guide updates to other model components? FedVGA explores this question through the interaction between classifiers and representations.

This interpretation does not establish equivalence to centralized optimization: multiple local updates and aggregation of locally fitted classifier directions can produce different optimization trajectories. The current evaluation is limited to supervised image classification with a common architecture and label space across clients. Each cycle also requires two client–server interaction stages, while the error-bound advantage of geometry-confidence weighting remains conditional on its association with local site error. These considerations motivate further study of the relationship to centralized updates and performance under matched communication budgets.

B Extended Related Work

B.1 Decoupled Representation and Classifier Learning

The interaction between representation learning and classifier optimization has been extensively studied in centralized learning, particularly under long-tailed and imbalanced data. A representative line of work shows that high-quality representations can still be learned from biased class distributions while the jointly optimized classifier remains substantially biased, motivating the decoupling of feature learning and classifier re-balancing (Kang et al., 2019; Cao et al., 2019; Zhou et al., 2020). These studies establish the broader principle that representation formation and decision-rule adaptation need not follow the same optimization process.

This principle has also motivated several federated learning methods. FedRep (Collins et al., 2021) learns a shared representation while maintaining client-specific prediction heads, primarily targeting personalized federated learning. FedBABU (Oh et al., 2021) freezes the classifier head during federated training and updates only the feature-extraction body, followed by head adaptation for personalization. FedRoD (Chen & Chao, 2021) explicitly separates generic and personalized predictors to reconcile global and personalized performance. Other studies focus more directly on classifier bias under heterogeneous data: CCVR (Luo et al., 2021) calibrates the classifier after federated training using virtual representations, while CReFF (Shang et al., 2022) re-trains the classifier using federated features under heterogeneous and long-tailed data.

More closely related to FedVGA, FedNH (Dai et al., 2023) and FedETF (Li et al., 2023b) use a shared classification head that remains fixed during local feature optimization. Both methods address representation learning under data heterogeneity and evaluate global as well as personalized performance. FedVGA shares their use of globally coordinated classifier supervision. Its distinction lies in how the shared classifier is refined and, in particular, how representation aggregation is incorporated into classifier adaptation.

Comparison with FedNH. FedNH (Dai et al., 2023) initializes class prototypes with a uniform configuration and progressively updates them using class-wise mean representations computed by locally trained feature extractors. The server combines the aggregated prototypes with the previous

global head through temporal smoothing and normalization. Thus, the local statistics used to update the shared classifier are obtained from client-specific feature mappings.

FedVGA introduces an intermediate representation aggregation step before classifier adaptation. After the *Align* stage, the server aggregates the updated feature extractors and redistributes the resulting global extractor. During *Adapt*, all participating clients hold this same extractor fixed and optimize their classifier directions through cross-entropy. The local directions are therefore estimated under a common feature mapping, and their updates explicitly optimize discrimination among competing classes rather than directly assigning class-wise feature means. GCA subsequently aggregates these directions using class-wise sample counts and directional coherence. The distinction from FedNH consequently involves both the representation model used to estimate the local classification structure and the rule used to update that structure.

Comparison with FedETF. FedETF (Li et al., 2023b) uses a synthetic simplex equiangular tight frame (ETF) as a shared classifier that remains fixed throughout federated training, and additionally supports post-training personalization. FedVGA instead allows the shared classifier directions to evolve across cycles. The key distinction is whether the shared decision geometry remains prescribed or is refined using distributed data during federated training. FedETF retains its predefined ETF classifier throughout this process, whereas FedVGA incorporates locally learned discriminative directions into an evolving global geometry. This allows the shared classification structure to adapt to the learned representations while preserving common supervision across clients.

FedVGA builds on shared classifier supervision and introduces a global–local–global refinement process that explicitly places representation aggregation before local classifier adaptation. The current global geometry $\mathcal{G}^t$ first coordinates representation learning across participating clients during *Align*. The updated feature extractors are then aggregated and redistributed, establishing a common feature mapping under which clients perform discriminative classifier adaptation during *Adapt*. GCA subsequently consolidates the locally adapted directions into $\mathcal{G}^{t+1}$, which provides shared supervision for the next cycle.

The defining feature of this process is the coordination between representation aggregation and geometry refinement. Local classifier directions are learned through a discriminative objective under the newly aggregated, frozen representation model, and their contributions to the next global geometry are modulated by class-wise statistical support and directional coherence. This differs from updating the shared head using feature means computed by client-specific extractors in FedNH, and from retaining a predefined classifier geometry throughout federated training in FedETF. Through this coupled refinement, FedVGA allows distributed discriminative information to reshape the global alignment target while keeping representation learning insulated from direct client-specific classifier adaptation within each cycle.

B.2 Adaptive Aggregation under Statistical Heterogeneity

Standard federated averaging assigns client contributions primarily according to local sample sizes. Beyond improving local optimization, another line of work revisits this aggregation rule and incorporates additional information about client reliability or distributional discrepancy. FedDisco (Ye et al., 2023), for example, complements dataset size with the discrepancy between local and global category distributions when determining aggregation weights. FedLAW (Li et al., 2023a) further studies the role of relative client importance and learns aggregation weights from federated training dynamics. Related optimization-based methods such as FedNova (Wang et al., 2020) and FedDyn (Acar et al., 2021) also modify the effect of heterogeneous local updates on the resulting global model through normalization or dynamic regularization.

GCA is related to this general principle of moving beyond purely sample-size-based aggregation, but differs in both the object being aggregated and the reliability signal being used. Rather than reweighting complete client models according to client-level distribution statistics or learned importance, GCA operates *class-wise* on locally adapted Voronoi sites. Its geometry-confidence factor is derived from the directional evidence of class representations in the aligned feature space, and therefore reflects how consistently local samples support the corresponding discriminative direction. The resulting weight $n_{k,c}(1 + \lambda\kappa_{k,c})$ retains class-wise statistical support while using geometry confidence to modulate the contribution of each local site. Accordingly, GCA serves specifically to

consolidate heterogeneous local decision geometries into the shared geometric target required by the next alignment cycle, rather than to redesign global model aggregation in general.

B.3 Voronoi-Based Federated Learning and Comparison with FedUD

FedUD (Lei et al., 2025) is closely related to FedVGA in that both exploit the Power/Voronoi interpretation of neural classifiers. Nevertheless, the two methods use this geometric view for substantially different purposes and consequently lead to different learning paradigms.

Problem setting and objective. FedUD is developed primarily for federated learning under feature/domain shift. It introduces a personalized generator for each client to transform local data toward a unified domain, while geometric constraints are incorporated to reduce inconsistencies among client decision structures. FedVGA instead studies cross-client representation alignment under broader statistical heterogeneity, including both label and feature shifts. Its objective is not to learn client-specific domain transformations, but to coordinate heterogeneous representations toward a common discriminative geometry shared by a single global model.

Role of Voronoi geometry. In FedUD, Power/Voronoi geometry provides geometric regularization for reducing domain and decision-boundary discrepancies during local optimization. The geometry therefore acts primarily as a constraint accompanying the task objective. FedVGA assigns a different role to classifier-induced geometry: the spherical Voronoi configuration itself constitutes the global alignment target. Its sites are treated jointly as a discriminative configuration, so the target encodes not only class-wise directions but also their relative decision structure. Representation learning is then performed directly against this common geometry rather than through an auxiliary geometric regularizer.

Optimization and geometric refinement. FedUD alternates the optimization of personalized generators and local models to progressively transform heterogeneous domains toward a unified one. FedVGA instead separates global representation alignment from local geometry adaptation. During *Align*, the global geometry is fixed and shared across all participating clients, preventing client-specific classifier adaptation from directly changing the supervision received by their representations. During *Adapt*, the aligned feature extractor is fixed and each client re-estimates its local classifier-induced geometry. Crucially, these local geometries are subsequently consolidated by GCA before they can affect feature learning in the next cycle. FedVGA therefore forms a global-geometry $\rightarrow$ representation $\rightarrow$ local-geometry $\rightarrow$ global-geometry refinement loop.

Overall, FedUD uses Power/Voronoi geometry together with personalized domain transformation to reduce feature-domain discrepancies, whereas FedVGA treats classifier-induced Voronoi geometry as the shared alignment object and progressively refines this object to coordinate representations across heterogeneous clients.

C Detailed Algorithm

We summarize the complete training procedure of FedVGA in Algorithm 1. Each communication cycle consists of an *Align* stage followed by an *Adapt* stage. The current global geometry first provides common supervision for local representation alignment. After the aligned feature extractors are aggregated, the updated global feature extractor is redistributed to participating clients, which then adapt their local classifier geometries with the representation space fixed. Finally, the resulting local geometries are consolidated by GCA to construct the shared geometry for the next cycle.

Notation. Let $\mathcal{S}^{t,\mathrm{Align}}$ and $\mathcal{S}^{t,\mathrm{Adapt}}$ denote the client subsets independently sampled for the *Align* and *Adapt* stages of cycle t, respectively. For class c, we further define $\mathcal{S}_c^{t,\mathrm{Adapt}} = \{k \in \mathcal{S}^{t,\mathrm{Adapt}} : n_{k,c} > 0\}$, which contains the participating Adapt clients with local samples from class c.

Algorithm 1 FEDVGA: Federated Voronoi Geometry Alignment

Input: Client Datasets $\{\mathcal{D}_k\}_{k=1}^{K}$; Number of Communication Cycles T; Geometry-confidence Coefficient λ.
Output: Global Feature Extractor $\mathcal{F}_{\theta^T}$ and Global Voronoi Geometry $\mathcal{G}^T = \{\boldsymbol{\mu}_c^T\}_{c=1}^{C}$.
Initialize: Global Feature Extractor $\mathcal{F}_{\theta^0}$ and Global Geometry $\mathcal{G}^0 = \{\boldsymbol{\mu}_c^0\}_{c=1}^{C}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Sample Participating Clients $\mathcal{S}^{t,\mathrm{Align}}$.
  Align: Voronoi Geometry-guided Representation Learning
  for each client $k \in \mathcal{S}^{t,\mathrm{Align}}$ **in parallel do**
    $\theta_k^t \leftarrow \theta^t$, $\mathcal{G}_k^t \leftarrow \mathcal{G}^t$ ▷ Initialize with global model
    Fix $\mathcal{G}_k^t$ and Update θ_k^t ▷ Using Eq. 3
    Upload θ_k^{t+1}
  end for
  $\theta^{t+1} \leftarrow \mathrm{FedAvg}(\{\theta_k^{t+1}\}_{k \in \mathcal{S}^{t,\mathrm{Align}}})$ ▷ Aggregate aligned representations
  Sample Participating Clients $\mathcal{S}^{t,\mathrm{Adapt}}$
  Adapt: Local Geometry Adaptation
  for each client $k \in \mathcal{S}^{t,\mathrm{Adapt}}$ **in parallel do**
    $\theta_k^{t+1} \leftarrow \theta^{t+1}$, $\mathcal{G}_k^t \leftarrow \mathcal{G}^t$ ▷ Initialize with global model
    Fix θ_k^{t+1} and Update $\mathcal{G}_k^t$ ▷ Using Eq. 4
    for each class c observed at client k **do**
      $n_{k,c} \leftarrow |\mathcal{D}_{k,c}|$
      Compute $\rho_{k,c}$ ▷ using Eq. 6
      Compute geometry-coherence score $\kappa_{k,c}$ ▷ using Eq. 7
    end for
    Upload $\{\boldsymbol{\mu}_{k,c}^{t+1}, \kappa_{k,c}, n_{k,c}\}_{c:n_{k,c}>0}$
  end for
  Geometry-Confidence Aggregation
  for each class $c = 1, \ldots, C$ **do**
    $\mathcal{S}_c^{t,\mathrm{Adapt}} \leftarrow \{k \in \mathcal{S}^{t,\mathrm{Adapt}} : n_{k,c} > 0\}$
    if $\mathcal{S}_c^{t,\mathrm{Adapt}} \neq \emptyset$ **then**
      for each $k \in \mathcal{S}_c^{t,\mathrm{Adapt}}$ **do**
        Compute $\alpha_{k,c}$ ▷ using Eq. 8
      end for
      $\widetilde{\boldsymbol{\mu}}_c^{t+1} \leftarrow \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \alpha_{k,c} \boldsymbol{\mu}_{k,c}^{t+1}$ ▷ using Eq. 9
      if $\|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2 > \varepsilon$ **then**
        $\boldsymbol{\mu}_c^{t+1} \leftarrow \widetilde{\boldsymbol{\mu}}_c^{t+1} / \|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2$
      else
        $\boldsymbol{\mu}_c^{t+1} \leftarrow \boldsymbol{\mu}_c^t$
      end if
    else
      $\boldsymbol{\mu}_c^{t+1} \leftarrow \boldsymbol{\mu}_c^t$
    end if
  end for
  $\mathcal{G}^{t+1} \leftarrow \{\boldsymbol{\mu}_c^{t+1}\}_{c=1}^{C}$ ▷ Global Voronoi geometry for next cycle
end for

D EXPERIMENTAL DETAILS

D.1 DATASET DETAILS

We conduct experiments on four image-classification benchmarks covering label and feature heterogeneity.

CIFAR-10 and CIFAR-100. CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) each contain $50{,}000$ training and $10{,}000$ test images with a resolution of 32×32. CIFAR-10 contains 10 classes, whereas CIFAR-100 contains 100 classes. To simulate label heterogeneity, only the training set is partitioned

Table 4: **Summary of the experimental configurations** for each dataset.

Configurations	CIFAR-10	CIFAR-100	Office-Caltech-10	DomainNet
Heterogeneity type	Label shift	Label shift	Feature shift	Feature shift
Number of clients	50 / 100 / 200	50 / 100 / 200	4	6
Input Resolution	$(32 \times 32 \times 3)$	$(32 \times 32 \times 3)$	$(128 \times 128 \times 3)$	$(128 \times 128 \times 3)$
Participation rate	0.5	0.5	1.0	1.0
Dirichlet parameter α	0.1 / 0.3 / 0.5	0.1 / 0.3 / 0.5	–	–
A&A cycles	100	100	100	100
Network	ResNet-18	ResNet-18	AlexNet	AlexNet
Batch size	128	128	64	64
Align local epochs	1	1	3	3
Adapt local epochs	3	3	3	3
Learning rate	0.1	0.1	0.01	0.01
Momentum	0.9	0.9	0.9	0.9
Weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
Scaling factor s	40	40	40	40
Geometry-confidence coefficient λ	0.4	0.4	0.8	0.8

across clients according to a Dirichlet distribution $\mathrm{Dir}(\alpha)$ over class proportions, following common FL protocols (Zhang et al., 2022; Acar et al., 2021). We consider $\alpha \in \{0.1, 0.3, 0.5\}$ and $K \in \{50, 100, 200\}$ clients, with 50% client participation in each training round. The original test set is retained as a common global test set for evaluating the learned global model.

Office-Caltech-10. Office-Caltech-10 (Gong et al., 2012) is a cross-domain recognition benchmark constructed from the 10 categories shared by Office-31 and Caltech-256. It consists of four visually distinct domains: *Amazon*, *Caltech*, *DSLR*, and *Webcam*. We treat each domain as one client, resulting in four clients with naturally different feature distributions.

DomainNet. DomainNet (Peng et al., 2019) is a large-scale multi-domain recognition benchmark containing six domains: *Clipart*, *Infograph*, *Painting*, *Quickdraw*, *Real*, and *Sketch*. For DomainNet, we select the first ten categories in alphabetical order and use the same category subset across all six domains. Each domain is treated as one client, and all methods use identical training and test splits.

For both feature-shift benchmarks, each domain is independently split into 80% training and 20% test samples. All domain clients participate in every communication cycle. A single global model is evaluated on each domain-specific test subset as well as their pooled test set.

D.2 Training Details

All methods are implemented in PyTorch and trained under the same data partitions, model architectures, and federated training protocol whenever applicable. We use ResNet-18 (He et al., 2016) for CIFAR-10/100 and AlexNet (Krizhevsky et al., 2012) for Office-Caltech-10 and DomainNet. The input images of CIFAR-10/100 are kept at 32×32, while images from Office-Caltech-10 and DomainNet are resized to 128×128. Standard data augmentation, including random cropping and horizontal flipping, is applied to the training sets, while only deterministic preprocessing is used for evaluation.

All models are randomly initialized. For FedVGA, the initial global geometry $\mathcal{G}^0 = \{\boldsymbol{\mu}_c^0\}_{c=1}^C$ is obtained by normalizing the randomly initialized classifier weights, i.e., $\boldsymbol{\mu}_c^0 = \mathbf{w}_c^0 / \|\mathbf{w}_c^0\|_2$.

For FedVGA, one reported round corresponds to one complete *Align-and-Adapt* cycle, consisting of two sequential client–server interaction stages. The *Align* and *Adapt* stages each use the same or fewer local epochs as the local update adopted by the competing methods.

For FedVGA, the hyperspherical scaling factor is fixed to $s = 40$ in all experiments. The geometry-confidence coefficient $\lambda \in [0, 1)$ is specified separately for each dataset and kept fixed across all heterogeneity settings of that dataset. The complete training configurations for all datasets are summarized in Table 4. Unless otherwise specified, all competing methods use the hyperparameters recommended in their original papers.

E Detailed Geometry Alignment and Boundary-Sensitive Analysis

This section provides detailed descriptions of the geometry-alignment and boundary-sensitive analyses presented in Fig. 1. These analyses characterize FedVGA from two complementary perspectives: the geometry-alignment analysis evaluates the consistency between local geometries and the shared global geometry during training, while the boundary-sensitive analysis investigates how the predictive improvements are distributed across samples with different levels of ambiguity.

E.1 Geometry Alignment Measurement

Global and Local Geometries. At communication cycle t, FedVGA starts with the global geometry $\mathcal{G}^t$ obtained from the previous communication cycle: $\mathcal{G}^t = \{\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_C^t\}$, where $\boldsymbol{\mu}_c^t \in \mathbb{S}^{d-1}$ denotes the global Voronoi site of class c.

Given $\mathcal{G}^t$, participating clients first update their feature extractors during Align. The server aggregates these extractors and redistributes the resulting global feature extractor for Adapt, during which participating clients hold it fixed and update their local classifier geometries. GCA then consolidates the adapted geometries $\mathcal{G}_k^{t+1}$ into $\mathcal{G}^{t+1}$. The alignment analysis compares these adapted local geometries with the aggregated global geometry within the same cycle.

Global-Local Geometry Alignment Score. After each communication cycle, we evaluate the consistency between each client's adapted local geometry and the corresponding aggregated global geometry. For client k and class c, the local-global geometric agreement is measured by the cosine similarity:

$$s_{k,c}^{t+1} = (\boldsymbol{\mu}_{k,c}^{t+1})^\top \boldsymbol{\mu}_c^{t+1}. \tag{14}$$

Since both geometry sites are normalized on the unit hypersphere, $s_{k,c}^{t+1} \in [-1, 1]$. A larger value indicates that the local decision geometry obtained after client-side adaptation is more consistent with the shared global geometry.

The overall Global-Local Geometry Alignment (GLGA) score is defined as:

$$\mathrm{GLGA}^{t+1} = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{|\mathcal{K}_c|} \sum_{k \in \mathcal{K}_c} (\boldsymbol{\mu}_{k,c}^{t+1})^\top \boldsymbol{\mu}_c^{t+1}, \tag{15}$$

where $\mathcal{K}_c$ denotes the set of clients whose local datasets contain class c, and C denotes the number of classes.

In centralized learning, heterogeneous data sources jointly contribute to a shared discriminative geometry without requiring their feature distributions or class means to coincide. FedVGA adopts this perspective by coordinating representation learning through a shared, progressively refined decision geometry. The local geometries evaluated by GLGA are obtained after Adapt, when each client fits its classifier to local data under the same frozen global feature extractor. These adapted geometries encode the class-discriminative structure learned from local representations. GLGA measures their agreement with the consolidated global geometry, assessing representation alignment through the consistency of locally learned decision structures with a common classification structure.

E.2 FedAvg+Proto Baseline

To ensure a fair comparison, FedAvg+Proto adopts the same model architecture and classifier parameterization as FedVGA. Therefore, its classifier parameters also induce a class-wise Voronoi decision geometry, which allows the same GLGA metric to be applied for both methods.

FedAvg+Proto follows the conventional FedAvg optimization paradigm, where the feature extractor and classifier are jointly optimized during local training. In addition, it introduces a prototype-

based regularization term to encourage local representations to approach empirical class-wise feature prototypes.

Specifically, for class c on client k, the empirical feature prototype is computed as:

$$\bar{\mathbf{z}}_{k,c} = \frac{1}{|\mathcal{D}_{k,c}|} \sum_{\mathbf{x} \in \mathcal{D}_{k,c}} F_{\theta_k}(\mathbf{x}). \tag{16}$$

The prototype regularization encourages representations belonging to the same class to approach their corresponding empirical feature prototypes.

After local optimization, the classifier parameters of FedAvg+Proto induce client-specific Voronoi geometries in the same manner as FedVGA. However, unlike FedVGA, these local geometries are obtained without the explicit Align-and-Adapt procedure. The resulting client geometries are aggregated using conventional sample-size weighted aggregation rather than Geometry-Confidence Aggregation:

$$\mathcal{G}^{t+1}_{\text{FedAvg+Proto}} = \text{Agg}_{\text{size}}(\mathcal{G}^{t+1}_1, \ldots, \mathcal{G}^{t+1}_K). \tag{17}$$

Therefore, the GLGA score of FedAvg+Proto is computed between its local classifier-induced geometries and the corresponding sample-size aggregated global geometry using the same evaluation protocol as FedVGA. This comparison isolates the effect of the proposed Align-and-Adapt procedure and Geometry-Confidence Aggregation from the general benefit of introducing class-level alignment objectives.

E.3 Boundary-Sensitive Sample Analysis

Prediction Margin. To characterize the ambiguity of samples under the baseline model, we use the prediction margin of FedAvg as a common reference. For a test sample $\mathbf{x}$, let $p^{(1)}_{\text{Avg}}(\mathbf{x})$ and $p^{(2)}_{\text{Avg}}(\mathbf{x})$ denote the largest and the second-largest predicted probabilities of FedAvg. The prediction margin is defined as

$$m(\mathbf{x}) = p^{(1)}_{\text{Avg}}(\mathbf{x}) - p^{(2)}_{\text{Avg}}(\mathbf{x}). \tag{18}$$

A smaller margin indicates that the top-ranked classes receive similar prediction confidence, corresponding to samples closer to the decision boundary and with higher prediction ambiguity.

Margin Distribution of Corrected Samples. We compare two types of samples according to their prediction outcomes under FedAvg and FedVGA. The FedAvg-correct set is defined as

$$\mathcal{S}_{\text{Avg}-\text{correct}} = \{(\mathbf{x}, y) : \hat{y}_{\text{Avg}}(\mathbf{x}) = y\}, \tag{19}$$

while the FedVGA-corrected set is defined as

$$\mathcal{S}_{\text{corrected}} = \{(\mathbf{x}, y) : \hat{y}_{\text{Avg}}(\mathbf{x}) \neq y, \hat{y}_{\text{VGA}}(\mathbf{x}) = y\}. \tag{20}$$

Figure 1(b) compares the distributions of $m(\mathbf{x})$ for these two subsets. The distributions are normalized independently and therefore describe the relative locations of the two sample groups along the margin axis rather than their absolute sample counts.

Margin-Conditioned Improvement. To quantify the performance difference under different ambiguity levels, we partition the test samples according to their FedAvg prediction margins:

$$\mathcal{B}_j = \{(\mathbf{x}, y) : m(\mathbf{x}) \in I_j\}, \tag{21}$$

where I_j denotes the j-th margin interval.

For each margin interval, we define the FedAvg-error, corrected, and regressed samples as

$$\mathcal{E}_j = \{(\mathbf{x}, y) \in \mathcal{B}_j : \hat{y}_{\mathrm{Avg}}(\mathbf{x}) \neq y\}, \tag{22}$$
$$\mathcal{C}_j = \{(\mathbf{x}, y) \in \mathcal{E}_j : \hat{y}_{\mathrm{VGA}}(\mathbf{x}) = y\}, \tag{23}$$
$$\mathcal{R}_j = \{(\mathbf{x}, y) \in \mathcal{B}_j : \hat{y}_{\mathrm{Avg}}(\mathbf{x}) = y, \hat{y}_{\mathrm{VGA}}(\mathbf{x}) \neq y\}. \tag{24}$$

The correction rate is defined as

$$\mathrm{CR}(I_j) = \frac{|\mathcal{C}_j|}{|\mathcal{E}_j|} \times 100\%. \tag{25}$$

The margin-conditioned accuracy gain is defined as

$$\Delta Acc(I_j) = \frac{|\mathcal{C}_j| - |\mathcal{R}_j|}{|\mathcal{B}_j|} \times 100\%. \tag{26}$$

Different from correction rate, which is conditioned on existing FedAvg errors, accuracy gain measures the net improvement over all samples within the margin interval and accounts for both corrected and regressed predictions.

The results in Fig. 1(b)-(c) provide complementary observations. The corrected samples are more concentrated in low-margin regions, and the largest accuracy gains are also observed in these regions. These results suggest that FedVGA is particularly effective for ambiguous samples that are difficult under the baseline model.

F Additional Theoretical Analysis

This section provides additional theoretical details and proofs for the results presented in the main text. We first briefly review the geometric error-propagation principle underlying the Simplex Lemma, which motivates the consideration of local geometric reliability in GCA. We then provide a detailed analysis of Align-and-Adapt (A&A), followed by the statistical interpretation and error-propagation analysis of Geometry-Confidence Aggregation (GCA).

Our analysis distinguishes between unconditional and conditional results. The results presented in the main text characterize the geometry-induced representation-gradient deviation and the propagation of local geometry errors without requiring any relationship between the proposed geometry confidence and the unknown local geometry error. We further derive a sufficient condition under which confidence-aware weighting yields a tighter error-propagation bound than conventional class-wise sample-size aggregation. Importantly, this condition is weak in the sense that it neither requires the geometry-coherence score to accurately estimate the local geometry error nor imposes a pointwise monotonic relationship between them. It only requires a non-positive aggregate association between directional coherence and local geometry error across clients. We therefore do not assume that GCA is universally superior to sample-size aggregation; rather, the conditional analysis provides a principled characterization of when the proposed confidence signal can translate into a tighter aggregation bound.

F.1 Geometric Background: Simplex Lemma

The design of GCA is motivated by a local-to-global error-propagation principle related to the Simplex Lemma of Ding and Xu (Ding & Xu, 2020). We briefly review the relevant geometric intuition here.

Consider a point set $Q \subset \mathbb{R}^d$ partitioned into J disjoint subsets, $Q = \bigcup_{j=1}^{J} Q_j, Q_i \cap Q_j = \emptyset, i \neq j$. Let $\mathbf{o}$ denote the mean of Q and $\mathbf{o}_j$ the mean of Q_j. Define $\delta^2 = \frac{1}{|Q|} \sum_{\mathbf{q} \in Q} \|\mathbf{q} - \mathbf{o}\|_2^2$.

The classical Simplex Lemma shows that the global mean can be approximated by searching within the simplex generated by the subset means. Of particular relevance here is its approximate form, where only inexact local representatives are available.

Lemma 1 (Approximate Simplex Lemma (Ding & Xu, 2020)). *Let $\{\mathbf{o}'_1, \dots, \mathbf{o}'_J\}$ be approximations of $\{\mathbf{o}_1, \dots, \mathbf{o}_J\}$ satisfying $\|\mathbf{o}'_j - \mathbf{o}_j\|_2 \le L, \qquad j = 1, \dots, J$. Let $\mathcal{V}'$ denote the simplex generated by $\{\mathbf{o}'_1, \dots, \mathbf{o}'_J\}$. For any $0 < \varepsilon \le 1$, there exists a point $\boldsymbol{\tau} \in \mathcal{V}'$ such that*

$$\|\boldsymbol{\tau} - \mathbf{o}\|_2 \le \sqrt{\varepsilon}\,\delta + (1+\varepsilon)L. \tag{27}$$

Lemma 1 reveals an important error-propagation property: the quality of a global geometric estimate depends explicitly on the accuracy of the local representatives. In particular, the local approximation error L directly enters the upper bound of the recovered global representative.

We emphasize that FedVGA does *not* directly apply Lemma 1. The lemma considers Euclidean subset means, whereas FedVGA communicates classifier-induced sites whose joint configuration defines a spherical Voronoi geometry. Instead, we borrow its underlying principle: when a global geometric structure is constructed from imperfect local representatives, reducing the contribution of unreliable local estimates can lead to a more reliable global target. This observation motivates the geometry-confidence weighting used by GCA.

F.2 Detailed Analysis of Align-and-Adapt

We first analyze the effect of local classifier-geometry adaptation on the representation-space gradient.

At communication cycle t, let $\mathcal{G}^t = \{\boldsymbol{\mu}_1^t, \dots, \boldsymbol{\mu}_C^t\}$ denote the shared global spherical Voronoi geometry. For a unit-normalized representation $\mathbf{z}$ with label y, the loss induced by an arbitrary geometry $\mathcal{G} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C\}$ is

$$\ell(\mathbf{z}, y; \mathcal{G}) = -\log \frac{\exp(s\boldsymbol{\mu}_y^\top \mathbf{z})}{\sum_{c=1}^{C} \exp(s\boldsymbol{\mu}_c^\top \mathbf{z})}, \tag{28}$$

where $s > 0$ is the scaling factor. Define the corresponding softmax probability as

$$p_c(\mathcal{G}) = \frac{\exp(s\boldsymbol{\mu}_c^\top \mathbf{z})}{\sum_{j=1}^{C} \exp(s\boldsymbol{\mu}_j^\top \mathbf{z})}. \tag{29}$$

The gradient with respect to the normalized representation is

$$g(\mathbf{z}, y; \mathcal{G}) = \nabla_{\mathbf{z}} \ell(\mathbf{z}, y; \mathcal{G}) = s\left(\sum_{c=1}^{C} p_c(\mathcal{G})\boldsymbol{\mu}_c - \boldsymbol{\mu}_y\right). \tag{30}$$

Under conventional joint feature–classifier optimization, let $\mathcal{G}_k^{t,e} = \{\boldsymbol{\mu}_{k,1}^{t,e}, \dots, \boldsymbol{\mu}_{k,C}^{t,e}\}$ denote the locally adapted classifier geometry of client k at local optimization step e. At the same representation state $z_{k,e}$, its representation gradient can be decomposed with respect to the shared geometry $\mathcal{G}^t$ as

$$\begin{aligned} g_{k,e}^{\text{joint}} &= g(\mathbf{z}_{k,e}, y; \mathcal{G}_k^{t,e}) \\ &= \underbrace{g(\mathbf{z}_{k,e}, y; \mathcal{G}^t)}_{g_{k,e}^{g}} + \underbrace{\left[g(\mathbf{z}_{k,e}, y; \mathcal{G}_k^{t,e}) - g(\mathbf{z}_{k,e}, y; \mathcal{G}^t)\right]}_{e_{k,e}}. \end{aligned} \tag{31}$$

Here $e_{k,e}$ isolates the change in geometric supervision caused by replacing the shared global geometry with the locally adapted geometry, while keeping the representation state fixed.

We next prove the geometry-induced representation-gradient deviation bound stated in Theorem 3 of the main text.

Proof of Theorem 3 (Geometry-Induced Representation-Gradient Deviation). Recall $\delta_{k,e} = \max_c \|\boldsymbol{\mu}_{k,c}^{t,e} - \boldsymbol{\mu}_c^t\|_2$. For unit-normalized representations and Voronoi sites, we show that

$$\|e_{k,e}\|_2 \le s(2+s)\delta_{k,e}. \tag{32}$$

Proof. For simplicity, denote $\boldsymbol{\mu}'_c = \boldsymbol{\mu}^{t,e}_{k,c}, \boldsymbol{\mu}_c = \boldsymbol{\mu}^t_c, \delta = \delta_{k,e}$, and let $p'_c = p_c(\mathcal{G}^{t,e}_k), p_c = p_c(\mathcal{G}^t)$. Using Eq. 30, we have

$$\begin{aligned}\frac{1}{s}e_{k,e} &= \sum_{c=1}^{C} p'_c\boldsymbol{\mu}'_c - \boldsymbol{\mu}'_y - \sum_{c=1}^{C} p_c\boldsymbol{\mu}_c + \boldsymbol{\mu}_y \\ &= \sum_{c=1}^{C}(p'_c - p_c)\boldsymbol{\mu}_c + \sum_{c=1}^{C} p'_c(\boldsymbol{\mu}'_c - \boldsymbol{\mu}_c) - (\boldsymbol{\mu}'_y - \boldsymbol{\mu}_y). \end{aligned} \tag{33}$$

Applying the triangle inequality and using $\|\boldsymbol{\mu}_c\|_2 = 1$, $\sum_c p'_c = 1$, and $\|\boldsymbol{\mu}'_c - \boldsymbol{\mu}_c\|_2 \le \delta$, we obtain

$$\begin{aligned}\frac{1}{s}\|e_{k,e}\|_2 &\le \left\|\sum_{c=1}^{C}(p'_c - p_c)\boldsymbol{\mu}_c\right\|_2 + \left\|\sum_{c=1}^{C} p'_c(\boldsymbol{\mu}'_c - \boldsymbol{\mu}_c)\right\|_2 + \|\boldsymbol{\mu}'_y - \boldsymbol{\mu}_y\|_2 \\ &\le \|p' - p\|_1 + 2\delta. \end{aligned} \tag{34}$$

It remains to bound the change in the softmax probabilities. Let $a_c = s\boldsymbol{\mu}_c^\top \mathbf{z}_{k,e}, a'_c = s\boldsymbol{\mu}_c'^\top \mathbf{z}_{k,e}$ denote the two logit vectors. Since $\|\mathbf{z}_{k,e}\|_2 = 1$,

$$|a'_c - a_c| = s|(\boldsymbol{\mu}'_c - \boldsymbol{\mu}_c)^\top \mathbf{z}_{k,e}| \le s\|\boldsymbol{\mu}'_c - \boldsymbol{\mu}_c\|_2 \le s\delta, \tag{35}$$

and therefore $\|a' - a\|_\infty \le s\delta$.

The Jacobian of the softmax mapping is $J_{\mathrm{sm}}(p) = \mathrm{diag}(p) - pp^\top$. For any vector h, $[J_{\mathrm{sm}}(p)h]_c = p_c\left(h_c - \sum_j p_j h_j\right)$. The induced norm of this Jacobian from ℓ_∞ to ℓ_1 satisfies

$$\|J_{\mathrm{sm}}(p)h\|_1 = \mathbb{E}_p|H - \mathbb{E}_p H| \le \sqrt{Var_p(H)} \le \|h\|_\infty. \tag{36}$$

Consequently, by integrating the Jacobian along the line segment between a and a',

$$\|p' - p\|_1 \le \|a' - a\|_\infty \le s\delta. \tag{37}$$

Substituting Eq. 37 into Eq. 34 yields $\|e_{k,e}\|_2 \le s(s\delta + 2\delta) = s(2+s)\delta$, which completes the proof. □

Implication for Align-and-Adapt. During the Align stage of FedVGA, the global geometry $\mathcal{G}^t$ is fixed across all participating clients. Hence,

$$\mathcal{G}^{t,e}_k = \mathcal{G}^t \quad \Longrightarrow \quad \delta_{k,e} = 0 \quad \Longrightarrow \quad e_{k,e} = 0. \tag{38}$$

Therefore,

$$g^{\mathrm{A\&A}}_{k,e} = g^g_{k,e} \tag{39}$$

with respect to the classifier-induced geometric supervision.

This result should be interpreted specifically. It does not imply that different clients receive identical gradients, since their samples and representation states can still differ, nor does it establish monotonic improvement of a representation-alignment metric or convergence of the overall non-convex federated optimization. Instead, it establishes that the additional representation-gradient deviation caused by client-specific classifier-geometry adaptation is absent during the Align stage. During the subsequent Adapt stage, the feature extractor is fixed, so the locally adapted geometries do not directly induce representation updates in the same communication cycle. They affect subsequent representation learning only after being consolidated into the next shared global geometry.

Effect of representation synchronization before Adapt. We next characterize how the feature mapping used for classifier adaptation affects the resulting local geometry. Fix a cycle t and a client k selected for *Adapt*. Compare the synchronized extractor $F^S = F_{\theta^{t+1}}$ with an alternative frozen local extractor F^L_k on this client's data. The latter can be the client's post-*Align* extractor when it participates in both stages; the comparison does not require the two participating client sets in Algorithm 1 to coincide. The two adaptation paths use the same initial raw classifier, minibatches,

and paired stochastic forward choices. Each extractor's parameters and running statistics are held fixed.

For the minibatch $\mathcal{B}_{k,e}$ used at step e, write $\mathbf{z}^L_{k,i,e}$ and $\mathbf{z}^S_{k,i,e}$ for the unit-normalized features of the same input under the two mappings. Define their discrepancy by

$$\Delta_k = \max_{0\le e<E} \frac{1}{|\mathcal{B}_{k,e}|} \sum_{i\in\mathcal{B}_{k,e}} \|\mathbf{z}^L_{k,i,e} - \mathbf{z}^S_{k,i,e}\|_2. \tag{40}$$

Let $H_{k,e}(W; F)$ be the gradient with respect to the raw classifier weights of the minibatch loss in Eq. 4, with directions $\boldsymbol{\mu}_c = \mathbf{w}_c/\|\mathbf{w}_c\|_2$. An identical coupled weight-decay term may be included in both paths.

Proposition 2 (Representation-Induced Geometry Adaptation Deviation)**.** *Consider E updates*

$$W^A_{k,e+1} = W^A_{k,e} - \eta H_{k,e}(W^A_{k,e}; F^A_k), \qquad A \in \{L, S\}, \tag{41}$$

where $F^S_k = F^S$, $\eta > 0$, and $W^L_{k,0} = W^S_{k,0}$. Assume that every classifier row in both paths, including their endpoints, has norm at least $a > 0$. For each fixed extractor and minibatch, assume

$$\|H_{k,e}(W; F) - H_{k,e}(W'; F)\|_F \le L_W \|W - W'\|_F \tag{42}$$

for the classifier pairs needed in the comparison, with $L_W \ge 0$. Set

$$M = \frac{s(s+2)}{a}, \qquad K_E = \eta M \sum_{j=0}^{E-1} (1 + \eta L_W)^j. \tag{43}$$

Then the final classifiers and their normalized directions satisfy

$$\|W^L_{k,E} - W^S_{k,E}\|_F \le K_E \Delta_k, \tag{44}$$

$$\max_c \|\boldsymbol{\mu}^L_{k,c} - \boldsymbol{\mu}^S_{k,c}\|_2 \le \frac{2K_E}{a} \Delta_k. \tag{45}$$

Proof. We first bound the change in a classifier gradient when only its input representation changes. At fixed W, for a sample with unit feature $\mathbf{z}$ and label y,

$$\nabla_{\mathbf{w}_c} \ell = \frac{s}{\|\mathbf{w}_c\|_2} (I - \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top)\big(p_c(W, \mathbf{z}) - \mathbf{1}_{\{y=c\}}\big)\mathbf{z}. \tag{46}$$

The row-wise linear map multiplying $s(p_c - \mathbf{1}_{\{y=c\}})\mathbf{z}$ has operator norm at most $1/a$. Let p, p' be the softmax vectors at unit features $\mathbf{z}, \mathbf{z}'$, respectively, and let $\mathbf{e}_y$ be the label's one-hot vector. The softmax bound in Eq. 37 implies $\|p - p'\|_1 \le s\|\mathbf{z} - \mathbf{z}'\|_2$. Therefore,

$$\begin{aligned} &\|(p - \mathbf{e}_y)\mathbf{z}^\top - (p' - \mathbf{e}_y)\mathbf{z'}^\top\|_F \\ &\quad \le \|p - p'\|_2 + \|p' - \mathbf{e}_y\|_2 \|\mathbf{z} - \mathbf{z}'\|_2 \\ &\quad \le (s+2)\|\mathbf{z} - \mathbf{z}'\|_2. \end{aligned} \tag{47}$$

Applying the row-wise maps in Eq. 46 and averaging over the paired minibatch gives

$$\|H_{k,e}(W; F^L_k) - H_{k,e}(W; F^S)\|_F \le M \Delta_k. \tag{48}$$

The weight-decay terms cancel in this comparison at fixed W.

Let $d_e = \|W^L_{k,e} - W^S_{k,e}\|_F$. Adding and subtracting $H_{k,e}(W^S_{k,e}; F^L_k)$ and using Eqs. 42 and 48 yields

$$d_{e+1} \le (1 + \eta L_W) d_e + \eta M \Delta_k. \tag{49}$$

Since $d_0 = 0$, unrolling the recursion proves Eq. 44. Finally, for any vectors with norms at least a,

$$\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|_2} - \frac{\mathbf{v}}{\|\mathbf{v}\|_2} \right\|_2 \le \frac{2}{a} \|\mathbf{u} - \mathbf{v}\|_2. \tag{50}$$

This follows by adding and subtracting $\mathbf{v}/\|\mathbf{u}\|_2$ and applying the reverse triangle inequality. Applying it row by row and using the Frobenius-norm bound proves Eq. 45. □

The same perturbation argument can be extended to momentum-based updates by augmenting the recursion with the momentum state.

Implication for the refinement loop. Proposition 2 quantifies the effect of the representation mapping on the geometry learned from a fixed client's data. Redistributing the aggregated extractor before *Adapt* establishes a common mapping for all local classifier fits. The resulting directions capture adaptation to heterogeneous data under this shared representation basis and are subsequently consolidated by GCA. The comparison isolates the contribution of the extractor choice; client data can still produce different fitted directions under the common extractor.

F.3 Detailed Analysis of Geometry-Confidence Aggregation

We next analyze GCA. The analysis proceeds in three steps. First, we establish the statistical interpretation of the geometry-coherence score $\kappa_{k,c}$. Second, we characterize how arbitrary convex aggregation weights propagate local geometry errors into the global estimate, which directly yields the result for GCA. Finally, we compare GCA with class-wise sample-size aggregation under an explicit sufficient condition.

For a fixed class c, let $\mathcal{S}_c^{t,\mathrm{Adapt}} = \left\{k \in \mathcal{S}^{t,\mathrm{Adapt}} : n_{k,c} > 0\right\}$ denote the participating clients containing class c.

Directional basis of geometry confidence. For client k and class c, let $\{\mathbf{z}_{k,i}\}_{i=1}^{n_{k,c}}$ denote the unit-normalized representations, and define

$$\bar{\mathbf{z}}_{k,c} = \frac{1}{n_{k,c}} \sum_{i=1}^{n_{k,c}} \mathbf{z}_{k,i}, \qquad \rho_{k,c} = \|\bar{\mathbf{z}}_{k,c}\|_2. \tag{51}$$

For $n_{k,c} > 1$, the geometry-coherence score used by GCA is

$$\kappa_{k,c} = \frac{n_{k,c}\rho_{k,c}^2 - 1}{n_{k,c} - 1}. \tag{52}$$

For completeness, we define the geometry-coherence score for the single-sample case. When $n_{k,c} > 1$, $\kappa_{k,c}$ is computed as the pairwise directional agreement defined in Eq. 52. When $n_{k,c} = 1$, pairwise directional information is unavailable. In this case, we set $\kappa_{k,c} = 0$. The directional-coherence analysis below considers $n_{k,c} > 1$, where the pairwise interpretation is well defined. For $n_{k,c} = 1$, we use the convention $\kappa_{k,c} = 0$; the subsequent aggregation and error-propagation results remain applicable under this definition.

We now prove Proposition 1 in the main text.

Proof of Proposition 1 (Directional Basis of Geometry Confidence). Since $\|\mathbf{z}_{k,i}\|_2 = 1$,

$$\begin{aligned}
\rho_{k,c}^2 &= \left\| \frac{1}{n_{k,c}} \sum_{i=1}^{n_{k,c}} \mathbf{z}_{k,i} \right\|_2^2 \\
&= \frac{1}{n_{k,c}^2} \sum_{i=1}^{n_{k,c}} \sum_{j=1}^{n_{k,c}} \mathbf{z}_{k,i}^\top \mathbf{z}_{k,j} \\
&= \frac{1}{n_{k,c}^2} \left(n_{k,c} + \sum_{i \neq j} \mathbf{z}_{k,i}^\top \mathbf{z}_{k,j} \right).
\end{aligned} \tag{53}$$

Therefore,

$$\kappa_{k,c} = \frac{1}{n_{k,c}(n_{k,c}-1)} \sum_{i \neq j} \mathbf{z}_{k,i}^\top \mathbf{z}_{k,j}. \tag{54}$$

Let $\mathbf{z}$ and $\mathbf{z}'$ be two independent draws from the class-c representation distribution at client k, and define $\tau_{k,c} = \mathbb{E}[\mathbf{z}^\top \mathbf{z}']$. Under the i.i.d. sampling condition in Proposition 1, every off-diagonal term in Eq. 54 has expectation $\tau_{k,c}$. Hence,

$$\begin{aligned}
\mathbb{E}[\kappa_{k,c}] &= \frac{1}{n_{k,c}(n_{k,c}-1)} \sum_{i \neq j} \mathbb{E}\left[\mathbf{z}_{k,i}^\top \mathbf{z}_{k,j}\right] \\
&= \tau_{k,c}.
\end{aligned} \tag{55}$$

This completes the proof.

Eq. 54 also gives a deterministic interpretation of $\kappa_{k,c}$: it is the average pairwise directional agreement between distinct class representations. Thus, $\kappa_{k,c}$ measures the directional evidence supporting a local class-wise geometry. Importantly, it does not directly estimate the unknown local site error considered below.

GCA converts this coherence score into

$$r_{k,c} = 1 + \lambda\kappa_{k,c}, \qquad 0 \le \lambda < 1. \tag{56}$$

Since each pairwise inner product lies in $[-1, 1]$, $\kappa_{k,c} \in [-1, 1]$, and therefore $r_{k,c} > 0$. The GCA weight is consequently well defined:

$$\alpha_{k,c} = \frac{n_{k,c} r_{k,c}}{\sum_{j \in \mathcal{S}_c^{t,\mathrm{Adapt}}} n_{j,c} r_{j,c}}. \tag{57}$$

Local-to-global geometry error propagation. Let $\boldsymbol{\mu}_c^\star$ be an arbitrary unit-norm reference site for class c. For each participating client, define its local geometry error as

$$\epsilon_{k,c} = \|\boldsymbol{\mu}_{k,c}^{t+1} - \boldsymbol{\mu}_c^\star\|_2. \tag{58}$$

No assumption is made here regarding how $\boldsymbol{\mu}_c^\star$ is generated or how $\epsilon_{k,c}$ is related to $\kappa_{k,c}$.

We first state the error-propagation property for general convex aggregation.

Lemma 2 (Convex Geometry Error Propagation)**.** *Let $\beta_{k,c} \ge 0$ satisfy $\sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \beta_{k,c} = 1$, and define*

$$\widetilde{\boldsymbol{\mu}}_c(\beta) = \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \beta_{k,c} \boldsymbol{\mu}_{k,c}^{t+1}. \tag{59}$$

Then,

$$\|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \le \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \beta_{k,c} \epsilon_{k,c}. \tag{60}$$

Proof. Since the weights sum to one,

$$\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star = \sum_k \beta_{k,c} \left(\boldsymbol{\mu}_{k,c}^{t+1} - \boldsymbol{\mu}_c^\star \right). \tag{61}$$

The triangle inequality immediately gives

$$\begin{aligned} \|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 &\le \sum_k \beta_{k,c} \left\| \boldsymbol{\mu}_{k,c}^{t+1} - \boldsymbol{\mu}_c^\star \right\|_2 \\ &= \sum_k \beta_{k,c} \epsilon_{k,c}, \end{aligned} \tag{62}$$

which completes the proof. □

Theorem 4 in the main text follows immediately by setting $\beta_{k,c} = \alpha_{k,c}$.

Proof of Theorem 4 (Local-to-Global Geometry Error Propagation). For GCA, $\widetilde{\boldsymbol{\mu}}_c^{t+1} = \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \alpha_{k,c} \boldsymbol{\mu}_{k,c}^{t+1}$, where $\alpha_{k,c} \ge 0$ and $\sum_k \alpha_{k,c} = 1$. Applying Lemma 2 yields

$$\left\| \widetilde{\boldsymbol{\mu}}_c^{t+1} - \boldsymbol{\mu}_c^\star \right\|_2 \le \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \alpha_{k,c} \epsilon_{k,c}, \tag{63}$$

which is exactly the result stated in Theorem 4.

This result requires no relationship between geometry coherence $\kappa_{k,c}$ and the unknown local error $\epsilon_{k,c}$. It characterizes how the chosen aggregation weights determine the contributions of local geometry errors to an upper bound on the pre-normalized global-site error.

Effect of hyperspherical normalization. FedVGA projects the aggregated site back onto the unit hypersphere. The following result characterizes how the preceding error bound transfers to the normalized site.

Define

$$B_c(\beta) = \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \beta_{k,c} \epsilon_{k,c}. \tag{64}$$

If $B_c(\beta) < 1$, then $\widetilde{\boldsymbol{\mu}}_c(\beta) \neq 0$ and the normalized site $\boldsymbol{\mu}_c(\beta) = \frac{\widetilde{\boldsymbol{\mu}}_c(\beta)}{\|\widetilde{\boldsymbol{\mu}}_c(\beta)\|_2}$ satisfies

$$\|\boldsymbol{\mu}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \leq 2B_c(\beta). \tag{65}$$

To see this, Lemma 2 and the reverse triangle inequality give

$$\|\widetilde{\boldsymbol{\mu}}_c(\beta)\|_2 \geq 1 - \|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \geq 1 - B_c(\beta) > 0. \tag{66}$$

Moreover,

$$\begin{aligned} \|\boldsymbol{\mu}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 &\leq \|\boldsymbol{\mu}_c(\beta) - \widetilde{\boldsymbol{\mu}}_c(\beta)\|_2 + \|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \\ &= |1 - \|\widetilde{\boldsymbol{\mu}}_c(\beta)\|_2| + \|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \\ &\leq 2\|\widetilde{\boldsymbol{\mu}}_c(\beta) - \boldsymbol{\mu}_c^\star\|_2 \\ &\leq 2B_c(\beta), \end{aligned} \tag{67}$$

where the second inequality again follows from the reverse triangle inequality.

This normalization result is supplementary to Theorem 4. The main-text result concerns the pre-normalized aggregate and therefore does not require the additional condition $B_c(\beta) < 1$.

Propagation to next-cycle representation gradients. The aggregated geometry supplies the shared supervision for the next *Align* stage. We now relate local site errors to the resulting representation-gradient deviation from a reference geometry.

Corollary 1 (Geometry-to-Supervision Error Propagation)**.** *Let $\mathcal{G}^\star = \{\boldsymbol{\mu}_c^\star\}_{c=1}^C$ be a unit-norm reference geometry in the same representation space. Suppose that every class has at least one contributing client and $\|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2 > \varepsilon$, where ε is the numerical threshold used by the retention rule. Thus, every site in $\mathcal{G}^{t+1}$ is obtained by normalizing its aggregate. Define*

$$B_{\max}^{t+1} = \max_{1 \leq c \leq C} B_c(\alpha) = \max_{1 \leq c \leq C} \sum_{k \in \mathcal{S}_c^{t,\mathrm{Adapt}}} \alpha_{k,c} \epsilon_{k,c}. \tag{68}$$

For any unit-normalized representation $\mathbf{z}$ and label y,

$$\left\|\mathbf{g}(\mathbf{z}, y; \mathcal{G}^{t+1}) - \mathbf{g}(\mathbf{z}, y; \mathcal{G}^\star)\right\|_2 \leq 2s(2+s)B_{\max}^{t+1}. \tag{69}$$

Proof. For each class, Theorem 4 and the reverse triangle inequality give

$$\begin{aligned} \left\|\boldsymbol{\mu}_c^{t+1} - \boldsymbol{\mu}_c^\star\right\|_2 &\leq \left|1 - \|\widetilde{\boldsymbol{\mu}}_c^{t+1}\|_2\right| + \left\|\widetilde{\boldsymbol{\mu}}_c^{t+1} - \boldsymbol{\mu}_c^\star\right\|_2 \\ &\leq 2\left\|\widetilde{\boldsymbol{\mu}}_c^{t+1} - \boldsymbol{\mu}_c^\star\right\|_2 \leq 2B_c(\alpha). \end{aligned} \tag{70}$$

Consequently, $\delta^{t+1} = \max_c \|\boldsymbol{\mu}_c^{t+1} - \boldsymbol{\mu}_c^\star\|_2 \leq 2B_{\max}^{t+1}$. Applying the geometry sensitivity bound of Theorem 3 to $\mathcal{G}^{t+1}$ and $\mathcal{G}^\star$ at the same representation $\mathbf{z}$ yields

$$\left\|\mathbf{g}(\mathbf{z}, y; \mathcal{G}^{t+1}) - \mathbf{g}(\mathbf{z}, y; \mathcal{G}^\star)\right\|_2 \leq s(2+s)\delta^{t+1} \leq 2s(2+s)B_{\max}^{t+1}. \tag{71}$$

□

This corollary connects geometry consolidation to subsequent representation learning: local site errors propagate through aggregation and normalization to the shared supervision used in the next *Align* stage. The comparison holds at the same representation state and requires no assumption relating geometry coherence to local site error.

The retention rule can be included by setting $d_c = \|\boldsymbol{\mu}_c^t - \boldsymbol{\mu}_c^\star\|_2$ for a retained class and $d_c = 2B_c(\alpha)$ for a class whose aggregate is normalized. The same argument then bounds the representation-gradient deviation by $s(2+s)\max_c d_c$.

Conditional comparison with sample-size aggregation. We finally characterize when the confidence factor used by GCA can yield a tighter error-propagation bound than conventional class-wise sample-size aggregation. This comparison is intentionally stated as a conditional result and is not required for the definition or application of GCA.

Define the class-wise sample-size distribution

$$q_{k,c} = \frac{n_{k,c}}{\sum_{j \in \mathcal{S}_c^{t,\mathrm{Adapt}}} n_{j,c}}, \qquad \sum_k q_{k,c} = 1. \tag{72}$$

We use $\mathbb{E}_{q_c}[\cdot]$ and $\mathrm{Cov}_{q_c}(\cdot,\cdot)$ to denote expectation and covariance over clients under this distribution.

Under conventional class-wise sample-size aggregation, the error-propagation bound is

$$B_c^{\mathrm{size}} = \sum_k q_{k,c}\epsilon_{k,c} = \mathbb{E}_{q_c}[\epsilon_{k,c}]. \tag{73}$$

Using $r_{k,c} = 1 + \lambda\kappa_{k,c}$, the GCA weight can be rewritten as $\alpha_{k,c} = \frac{q_{k,c}r_{k,c}}{\mathbb{E}_{q_c}[r_{k,c}]}$, and its corresponding error-propagation bound is

$$B_c^{\mathrm{GCA}} = \frac{\mathbb{E}_{q_c}[r_{k,c}\epsilon_{k,c}]}{\mathbb{E}_{q_c}[r_{k,c}]}. \tag{74}$$

The two bounds admit the following exact relationship.

Proposition 3 (Conditional Comparison with Sample-Size Aggregation)**.** *For every class c,*

$$B_c^{\mathrm{GCA}} - B_c^{\mathrm{size}} = \frac{\mathrm{Cov}_{q_c}(r_{k,c},\epsilon_{k,c})}{\mathbb{E}_{q_c}[r_{k,c}]} = \frac{\lambda\ \mathrm{Cov}_{q_c}(\kappa_{k,c},\epsilon_{k,c})}{1+\lambda\mathbb{E}_{q_c}[\kappa_{k,c}]}. \tag{75}$$

Consequently, if

$$\mathrm{Cov}_{q_c}(\kappa_{k,c},\epsilon_{k,c}) \le 0, \tag{76}$$

then

$$B_c^{\mathrm{GCA}} \le B_c^{\mathrm{size}}. \tag{77}$$

Proof. By the definition of covariance,

$$\mathbb{E}_{q_c}[r_{k,c}\epsilon_{k,c}] = \mathbb{E}_{q_c}[r_{k,c}]\mathbb{E}_{q_c}[\epsilon_{k,c}] + \mathrm{Cov}_{q_c}(r_{k,c},\epsilon_{k,c}). \tag{78}$$

Substituting this identity into Eq. 74 gives

$$\begin{aligned} B_c^{\mathrm{GCA}} &= \mathbb{E}_{q_c}[\epsilon_{k,c}] + \frac{\mathrm{Cov}_{q_c}(r_{k,c},\epsilon_{k,c})}{\mathbb{E}_{q_c}[r_{k,c}]} \\ &= B_c^{\mathrm{size}} + \frac{\mathrm{Cov}_{q_c}(r_{k,c},\epsilon_{k,c})}{\mathbb{E}_{q_c}[r_{k,c}]}. \end{aligned} \tag{79}$$

Because $r_{k,c} = 1 + \lambda\kappa_{k,c}$,

$$\mathrm{Cov}_{q_c}(r_{k,c},\epsilon_{k,c}) = \lambda\,\mathrm{Cov}_{q_c}(\kappa_{k,c},\epsilon_{k,c}), \tag{80}$$

$$\mathbb{E}_{q_c}[r_{k,c}] = 1 + \lambda\mathbb{E}_{q_c}[\kappa_{k,c}] > 0. \tag{81}$$

Eq. 75 follows directly. Therefore, Eq. 76 implies Eq. 77. □

Interpretation of the sufficient condition. Condition 76 states that, under the class-wise sample-size distribution, clients with higher directional coherence tend to have no larger local geometry errors. Under this condition, confidence-aware reweighting yields a no-larger error-propagation upper bound than sample-size weighting.

Importantly, FedVGA does not assume Eq. 76 in its optimization procedure, and the unconditional results in Proposition 1 and Theorem 4 do not depend on it. The condition is introduced only to characterize when the confidence signal provides an additional advantage over sample-size aggregation. Nor do we claim that $\kappa_{k,c}$ is an estimator of the unknown $\epsilon_{k,c}$: Proposition 1 establishes only that it measures class-wise directional agreement. Because $\mu_c^\star$ is an analytical reference rather than

an observable ground-truth site, Eq. 76 is intended as a sufficient theoretical characterization rather than an empirically verifiable assumption required by FedVGA.

When $\lambda = 0$, $r_{k,c} = 1$ and GCA exactly reduces to class-wise sample-size aggregation:

$$\alpha_{k,c} = q_{k,c}, \qquad B_c^{\mathrm{GCA}} = B_c^{\mathrm{size}}. \tag{82}$$

For $\lambda > 0$, GCA changes the relative aggregation mass according to the observed directional coherence. Whether this reweighting tightens the error-propagation bound relative to sample-size aggregation is precisely characterized by Eq. 75.

Finally, if $B_c^{\mathrm{size}} < 1$ and Condition 76 holds, then $B_c^{\mathrm{GCA}} \le B_c^{\mathrm{size}} < 1$. Applying Eq. 65 to the two aggregation rules gives the corresponding normalization-aware upper bounds

$$2B_c^{\mathrm{GCA}} \le 2B_c^{\mathrm{size}}. \tag{83}$$

This comparison concerns the derived worst-case upper bounds; it does not assert that the realized normalized-site error produced by GCA must be smaller for every possible collection of local geometries.

G Detailed Ablation of the Align-and-Adapt Loop

We examine the roles of intermediate feature aggregation, geometry adaptation, and alternating optimization within the Align-and-Adapt loop. We compare full FedVGA with three variants on CIFAR-100, DomainNet, and Office-Caltech-10 in Table 5. Each variant produces a single global model for evaluation.

Variants.

- **Deferred aggregation.** Each client first performs *Align* under the shared global geometry and then performs *Adapt* using its own locally updated feature extractor, which remains frozen during classifier adaptation. The server aggregates the feature extractors only after both local stages have finished. Geometry-confidence scores are subsequently computed using the aggregated global extractor, and the adapted classifier directions are consolidated through GCA. The resulting global extractor and aggregated geometry jointly form the evaluation model. This variant examines the role of establishing a common representation mapping before local geometry adaptation.
- **Fixed orthogonal geometry.** The classifier directions are initialized as mutually orthogonal unit vectors and remain fixed throughout training. Clients update their feature extractors under this shared geometry, and the server aggregates the updated extractors. The global model is evaluated using the aggregated extractor and the fixed classifier. This variant examines whether adapting the shared geometry provides benefits beyond a well-separated, fixed geometric target.
- **Joint optimization (w/o A&A).** This is the same w/o A&A variant evaluated in the main text. Each client jointly optimizes its feature extractor and classifier on local data. The server aggregates feature extractors using client sample-size weights and consolidates the adapted classifier directions through GCA. The resulting global feature extractor and classifier constitute the model used for evaluation. This variant examines the effect of removing the separation between representation learning and local geometry adaptation.

Table 5: **Detailed ablation of the Align-and-Adapt loop.** Test accuracy (%) is evaluated on CIFAR-100 with $(K, \alpha) = (100, 0.1)$, the DomainNet 10-class subset, and Office-Caltech-10.

Variant	CIFAR-100	DomainNet	Office-Caltech-10
Deferred aggregation	28.39	60.45	59.84
Fixed orthogonal geometry	31.27	61.88	71.65
Joint optimization (w/o A&A)	27.59	55.32	61.61
Full FedVGA	37.93	66.02	77.16

Table 6: **Computational cost and extended-training comparison** under feature heterogeneity. DomainNet denotes the 10-class subset used in the main experiments. $\bar{t}$ is the average computational time per baseline round or complete FedVGA cycle, measured in seconds. Acc.@200R reports baseline accuracy after 200 rounds; Acc.@100C reports FedVGA accuracy after 100 cycles. Accuracies are expressed as percentages. "–" denotes a non-applicable entry.

Method	**DomainNet**			**Office-Caltech-10**		
	$\bar{t}$ (s)	Acc.@200R	Acc.@100C	$\bar{t}$ (s)	Acc.@200R	Acc.@100C
FedAvg	11.80 ± 0.54	53.40	–	22.94 ± 0.24	64.96	–
FedProx	18.38 ± 0.99	55.76	–	33.40 ± 0.27	61.84	–
FedNH	21.18 ± 0.49	57.28	–	27.44 ± 0.83	72.11	–
FedETF	32.16 ± 0.42	62.22	–	31.06 ± 0.05	73.70	–
FedVGA	19.85 ± 0.84	–	66.52	31.88 ± 0.89	–	76.99

H Computational Cost and Extended-Training Comparison

We complement the main accuracy comparisons with an analysis of computational cost and baseline performance under extended training. Each baseline is trained for 200 communication rounds, while FedVGA is trained for 100 complete Align-and-Adapt cycles. Since each FedVGA cycle contains two sequential client–server interaction stages, these settings match the number of interaction stages. They do not necessarily match the total computation or the volume of transmitted data.

Experimental protocol. All methods use the same data partitions, backbone architectures, and preprocessing within each dataset. We retain the method-specific local training configurations used in the main experiments and extend the baseline training horizon to 200 rounds.

Computational cost. Timing is performed under the same hardware configuration and client execution policy. For a baseline, one iteration denotes one communication round; for FedVGA, one iteration denotes a complete Align-and-Adapt cycle. The measured computational time includes local training, method-specific auxiliary computations, and server aggregation. In particular, the FedVGA measurement includes both Align and Adapt, geometry-coherence estimation, and GCA. Test evaluation, checkpoint writing, and network transmission latency are excluded. Thus, the reported time characterizes computational cost in the experimental environment.

Let $t_{m,r}$ denote the measured time of iteration r for method m. We report the average iteration time

$$\bar{t}_m = \frac{1}{R_m} \sum_{r=1}^{R_m} t_{m,r}, \qquad R_m = \begin{cases} 200, & \text{for a baseline,} \\ 100, & \text{for FedVGA.} \end{cases} \tag{84}$$

Evaluation settings. Table 6 reports results on the DomainNet 10-class subset and Office-Caltech-10 under feature heterogeneity. Table 7 reports results on CIFAR-10 and CIFAR-100 under label heterogeneity, both using $(K, \alpha) = (100, 0.1)$ and a client participation rate of 0.5. For each dataset, we report average iteration time, baseline accuracy after 200 rounds, and FedVGA accuracy after 100 cycles. An entry marked "–" indicates that the corresponding training horizon is not applicable to that method.

Table 7: **Computational cost and extended-training comparison** under label heterogeneity on CIFAR-10 and CIFAR-100. Both datasets use $(K, \alpha) = (100, 0.1)$ with 50% client participation per interaction stage. $\bar{t}$ is the average computational time per baseline round or complete FedVGA cycle, measured in seconds. Acc.@200R reports baseline accuracy after 200 rounds; Acc.@100C reports FedVGA accuracy after 100 cycles. Accuracies are expressed as percentages. "–" denotes a non-applicable entry.

Method	**CIFAR-10**			**CIFAR-100**		
	$\bar{t}$ (s)	Acc.@200R	Acc.@100C	$\bar{t}$ (s)	Acc.@200R	Acc.@100C
FedAvg	21.51 ± 0.09	45.31	–	22.23 ± 0.28	28.84	–
FedProx	30.29 ± 0.12	48.78	–	30.66 ± 0.08	26.74	–
FedNH	28.10 ± 0.26	58.38	–	23.06 ± 0.32	31.26	–
FedETF	26.41 ± 0.88	56.63	–	21.49 ± 0.19	28.01	–
FedVGA	26.25 ± 0.09	–	60.58	39.44 ± 0.43	–	38.01