FedQ-Vision: Alignment-Free Federated Visual Learning via Quotient Geometry

Tirtho Roy Ushashi Bhattacharjee Koushik Howlader
Iowa State University
tirtho@iastate.edu

Abstract

Federated visual learning increasingly involves clients running different *frozen vision foundation models whose representation coordinate systems are neither comparable nor shareable: parameters cannot be averaged, and existing methods must learn cross-space alignment maps, usually from shared pilot data. We introduce FedQ-Vision, an alignment-free framework in which each client transmits only a compressed, coordinate-invariant summary—the Gram matrix of its class prototypes—from which the server recovers a global visual* quotient geometry *in a single round, with no alignment and no representation training, backed by explicit recovery guarantees (invariance, identifiability, optimal low-rank communication, and neighbourhood stability). As a flagship aggregator we introduce a* learnable-gate *rule that fuses this quotient geometry with a whitened shared-anchor representation,* learning *per class-pair how far to trust each, to remain robust under extreme label skew. Across six standard benchmarks—spanning handwritten digit/character, apparel/object, and* 200*-class recognition—under heterogeneous CNN and transformer encoders, FedQ-Vision attains the best test-image classification accuracy of all methods: a statistically significant* $+9.4$*-point gain over the no-collaboration baseline, surpassing all six federated baselines on every dataset, at one to two orders of magnitude lower communication in a single round. Requiring only frozen encoders and class prototypes, it extends to new encoders without retraining and improves as foundation models do. Code:* https://github.com/tirtho149/FedQ-Vision.

1. Introduction

Federated visual learning increasingly involves clients that run *different* frozen vision foundation models (DINOv2, CLIP, SigLIP, MAE) [10, 22, 23, 34]. Such clients cannot average parameters (architectures and feature dimensions differ), and their latent spaces are not directly comparable. The dominant remedy is *representation alignment*: learn maps between client latent spaces [4, 21], typically requiring shared pilot samples and iterative optimization.

We ask a sharper question: *is alignment necessary at all?* Our starting point is that the task-relevant structure of a set of class prototypes lives not in their absolute coordinates but in their *pairwise geometry*, which is invariant to the orthogonal gauge relating different encoders' coordinate systems (Fig. 1). If the server only needs this geometry, clients can transmit a coordinate-invariant summary—the prototype Gram matrix—and the server can aggregate summaries directly, with no alignment maps, no shared pilot for alignment, and no iterative training.

FedQ-Vision makes this principle a practical algorithm. Each client (i) freezes its encoder, (ii) forms one prototype per class, (iii) builds the invariant Gram geometry, (iv) compresses it to rank r, and (v) transmits only the compressed factors and support counts. The server reliability-weights the summaries into a global quotient geometry usable for retrieval, nearest-neighbour transfer and clustering. We contribute:

- **Theory.** Exact orthogonal invariance and quotient identifiability of the prototype Gram (Theorems 1 and 2), an unbiased-averaging concentration bound (Theorem 3), the optimal low-rank communication trade-off (Theorem 4), nearest-neighbour preservation (Theorem 5), and graceful degradation under approximate orthogonality (Theorem 6).
- **Algorithm.** A one-shot, alignment-free aggregator, plus a support-adaptive variant that fuses direct quotient geometry with a *whitened shared-anchor* (Nyström) representation [21, 25, 27] to remain robust under extreme non-IID skew.
- **Evaluation.** On six datasets drawn from the federated baselines' own protocols, under truly heterogeneous frozen encoders, FedQ-Vision matches or beats every applicable baseline on downstream utility and geometry recovery at 30–300$\times$ lower communication and a single round, with 5-seed paired significance tests.

2. Related Work

Parameter-averaging FL. FedAvg [20] and FedProx [18] average model weights and therefore require architecturally

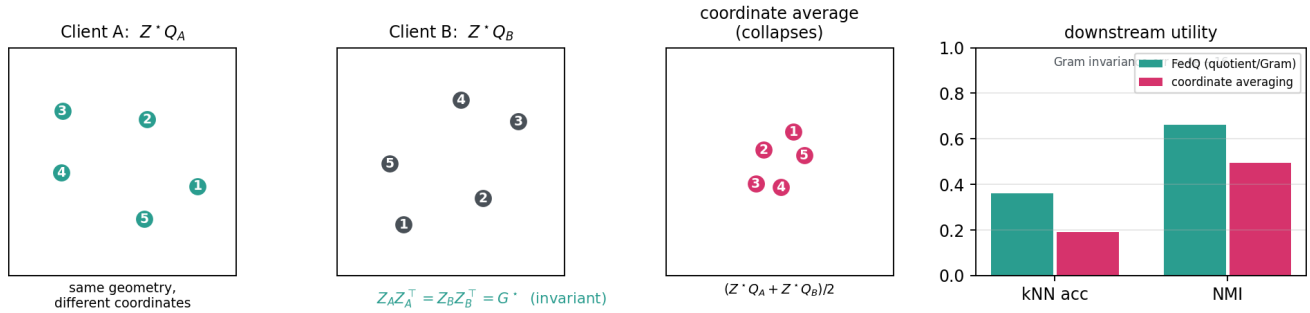

Figure 1. **The central idea.** The same class configuration $Z^\star$ seen in two clients' coordinate frames, $Z^\star Q_A$ and $Z^\star Q_B$, looks different coordinate-wise, yet the Gram matrix is identical: $Z_A Z_A^\top = Z_B Z_B^\top = G^\star$ (Thm. 1). A coordinate-based federated method that averages coordinates *collapses* the geometry, whereas FedQ-Vision aggregates the invariant Gram and is unaffected. **Right:** in a controlled experiment that scrambles every client's coordinates with an independent orthogonal gauge, quotient (Gram) aggregation retains full downstream utility while coordinate averaging collapses; the Gram is invariant to floating-point precision.

compatible clients; they are undefined across heterogeneous frozen encoders and serve only as a homogeneous-track reference.

Model-heterogeneous / prototype FL. Surveys [8, 32] chart methods that share abstract knowledge instead of weights. FedProto [26] and FedTGP [35] exchange class prototypes, but averaging prototypes implicitly assumes a *shared coordinate system*: under different encoders the coordinates are incomparable and the aggregate collapses (Sec. 6).

Relation- and topology-level sharing. Relation-Level FedTopo [19] broadcasts a class-relation topology to guide multi-round local training, and FedTopo [11] adds a topological alignment loss. These are the closest conceptual comparators; FedQ-Vision differs by being *one-shot*, *compressed*, and requiring *no local training*.

Representation alignment and relative representations. Sheaf-FRL [4] learns orthogonal/Stiefel restriction maps from a shared pilot set via alternating Procrustes [24] updates. Relative representations [21] encode each sample by its similarities to fixed anchors, an isometry-invariant code; learned/whitened anchors improve them [25]. FedQ-Vision connects this line to federated geometry: our whitened-anchor component is a Nyström [27] feature map used *only* as a skew-robust fallback, never to align coordinates.

Anchor-based and orthogonal FL. A growing line uses anchors, confidence weighting or structural constraints to mitigate heterogeneity: FedSA [36] learns semantic anchors as prototypes, RefProtoFL [28] uses reference prototypes, CAFedCL [29] adds confidence-aware weighting, FedDAP [15] is domain-aware, and FedOT [12] personalizes black-box foundation models with local orthogonal transforms. We instead use shared anchors only as a skew-robust read-out and never learn orthogonal or cross-space maps. Recent work also rethinks prototype alignment as structural rather than coordinate matching [30, 33], echoing our quotient-geometry stance.

Table 1. Qualitative comparison of federated methods. ✓: supported / required-free; ✗: not; ∼: limited. "No repr. train": no local representation/encoder training (FedQ-Vision only fits two scalar gate parameters on held-out validation, not a model). Collab.: enables cross-client knowledge sharing (Local-Only does not). FedQ-Vision is the only method that is model-heterogeneous, alignment-free, free of representation training, *and* one-shot while still collaborating.

Method	Venue	Model heterog.	Align. free	No repr. train	One shot	Collab.
Local-Only	–	✓	✓	✓	✓	✗
FedAvg [20]	AISTATS'17	✗	✗	✗	✗	✓
FedProx [18]	MLSys'20	✗	✗	✗	✗	✓
FedProto [26]	AAAI'22	✓	✗	✗	✗	✓
FedTopo [11]	AAAI'26	∼	✗	✗	✗	✓
Rel.-FedTopo [19]	arXiv'26	✓	✓	✗	✗	✓
Sheaf-FRL [4]	arXiv'26	✓	✗	✗	✗	✓
FedQ-Vision (Ours)	Ours	✓	✓	✓	✓	✓

One-shot statistics and universe alignment. One-shot federated ridge regression aggregates client Gram/moment sufficient statistics once to recover the centralized solution [2]; FedQ-Vision shares the "transmit a coordinate-invariant second-order summary once" principle but targets *cross-encoder* quotient geometry with no compatible coordinates. Generalized-Procrustes/universe methods [1] instead build a shared orthogonal reference by *learning* alignments across spaces, which FedQ-Vision avoids entirely.

Foundation-model geometry. Different encoders exhibit high representational similarity under CKA [13]; we quantify this for our encoders (Sec. 5) and exploit it empirically, while our theory rests only on the weaker quotient-geometry assumption.

Table 1 summarizes the qualitative distinctions. FedQ-Vision is the only method that is simultaneously model-

heterogeneous, alignment-free, training-free, and one-shot while still enabling collaboration (unlike Local-Only).

3. Problem Setup and Theory

Consider K clients; client k owns private images D_k and a frozen encoder f_k, producing $Z_k = f_k(D_k) \in \mathbb{R}^{m \times d_k}$. Encoders may differ in architecture and dimension d_k. For the clean model, corresponding visual states share a canonical configuration $Z^\star \in \mathbb{R}^{m \times r}$ of intrinsic dimension $r \le \min_k d_k$, and each client embeds it isometrically into its own representation space,

$$Z_k = Z^\star Q_k, \qquad Q_k \in \mathbb{R}^{r \times d_k}, \quad Q_k Q_k^\top = I_r, \qquad (1)$$

i.e. Q_k is a *semi-orthogonal* map (orthonormal rows) that can embed the same r-dimensional configuration into spaces of *different* dimensions d_k. The target is not $Z^\star$ but its equivalence class $[Z^\star] = \{Z^\star U : U \in \mathrm{O}(r)\}$; the server seeks the quotient object $G^\star = Z^\star Z^{\star\top}$.

Theorem 1 (Invariance under isometric client embeddings)**.** *If $Z_k = Z^\star Q_k$ with $Q_k Q_k^\top = I_r$ (Eq. (1)), then $G_k = Z_k Z_k^\top = G^\star$, independently of d_k and Q_k.*

Proof. Using $Q_k Q_k^\top = I_r$, $G_k = (Z^\star Q_k)(Z^\star Q_k)^\top = Z^\star Q_k Q_k^\top Z^{\star\top} = Z^\star I_r Z^{\star\top} = Z^\star Z^{\star\top} = G^\star$. Both the client's dimension and its embedding cancel exactly, so a client with a 2048-d and one with a 768-d encoder produce the same $C \times C$ summary. □

Theorem 2 (Quotient identifiability)**.** *Let $X \in \mathbb{R}^{m \times d_1}$, $Y \in \mathbb{R}^{m \times d_2}$ each have rank r. If $XX^\top = YY^\top$ then $Y = XW$ for a semi-orthogonal W, and both share the same r-dimensional configuration up to $\mathrm{O}(r)$. Absolute coordinates are non-identifiable; the quotient geometry is identifiable.*

Proof. Let $G = XX^\top = YY^\top$ have rank r, with eigendecomposition $G = U\Lambda U^\top$, $\Lambda \succ 0$ on its r-dimensional support. Any factor F ($m \times d$, rank r) with $FF^\top = G$ satisfies $F = U\Lambda^{1/2}R^\top$ for a semi-orthogonal $R \in \mathbb{R}^{d \times r}$ ($R^\top R = I_r$). Hence $X = U\Lambda^{1/2}R_X^\top$ and $Y = U\Lambda^{1/2}R_Y^\top$ share the same intrinsic configuration $U\Lambda^{1/2}$ up to the $\mathrm{O}(r)$ ambiguity in R_X, R_Y, and $Y = XW$ with $W = R_X R_Y^\top$; the Gram G is invariant to this ambiguity. □

Corollary 1. *$\|z_i - z_j\|_2^2 = G_{ii} + G_{jj} - 2G_{ij}$: recovering G recovers all pairwise Euclidean distances.*

Proof. $\|z_i - z_j\|_2^2 = \langle z_i, z_i\rangle + \langle z_j, z_j\rangle - 2\langle z_i, z_j\rangle = G_{ii} + G_{jj} - 2G_{ij}$. □

Theorem 3 (Concentration under unbiased client geometry noise)**.** *Adopt the conditional statistical model in which client k sends $G_k = G^\star + E_k$ with $\mathbb{E}[E_k]=0$, the E_k independent, and $\mathbb{E}\|E_k\|_F^2 \le \sigma_k^2/n_k$. Then for weights $w_k \ge 0$, $\sum_k w_k=1$, $\hat{G} = \sum_k w_k G_k$ satisfies $\mathbb{E}\|\hat{G} - G^\star\|_F^2 \le \sum_k w_k^2 \sigma_k^2/n_k$, i.e. $\sigma^2/(Kn)$ for balanced clients.*

Remark. This is a statement about a client-geometry noise model, not about the raw estimator: prototype estimation, row-normalization and the Gram outer product are nonlinear, and cross-encoder mismatch may be systematic rather than zero-mean, so the model holds only to the extent that the residual E_k is approximately unbiased and independent (Theorem 6 bounds the deterministic part of the error separately).

Proof. With $\hat{G} - G^\star = \sum_k w_k E_k$, expanding the squared Frobenius norm gives $\sum_k w_k^2 \|E_k\|_F^2 + \sum_{k \ne l} w_k w_l \langle E_k, E_l\rangle_F$. Independence and $\mathbb{E}[E_k] = 0$ make every cross term vanish in expectation, leaving $\mathbb{E}\|\hat{G} - G^\star\|_F^2 = \sum_k w_k^2 \mathbb{E}\|E_k\|_F^2 \le \sum_k w_k^2 \sigma_k^2/n_k$; balanced clients give $\sigma^2/(Kn)$. □

Theorem 4 (Optimal low-rank communication)**.** *For PSD G with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_m$, the rank-$r$ truncation G_r minimizes $\|G - A\|_F$ over rank-r A, with $\|G - G_r\|_F^2 = \sum_{j>r} \lambda_j^2$ (Eckart–Young [7]).*

Proof. G is symmetric PSD, so its eigendecomposition equals its SVD with singular values λ_j. By Eckart–Young–Mirsky [7], over all rank-r A the truncated eigendecomposition $G_r = \sum_{j \le r} \lambda_j u_j u_j^\top$ minimizes $\|G - A\|_F$, with residual $\|G - G_r\|_F^2 = \sum_{j>r} \lambda_j^2$. □

Theorem 5 (Nearest-neighbour preservation)**.** *Let $\hat{G} = G + \Delta$. If j is the true nearest neighbour of i and every competitor l has $d_{il}^2 - d_{ij}^2 > \gamma$, then j remains the nearest neighbour whenever $8\|\Delta\|_2 < \gamma$.*

Proof. By Corollary 1, $\hat{d}_{ab}^2 - d_{ab}^2 = \Delta_{aa} + \Delta_{bb} - 2\Delta_{ab}$, and each entry obeys $|\Delta_{ab}| = |e_a^\top \Delta e_b| \le \|\Delta\|_2$, so $|\hat{d}_{ab}^2 - d_{ab}^2| \le 4\|\Delta\|_2$. Hence for any competitor l, $\hat{d}_{il}^2 - \hat{d}_{ij}^2 \ge (d_{il}^2 - d_{ij}^2) - 8\|\Delta\|_2 \ge \gamma - 8\|\Delta\|_2 > 0$, so j stays the nearest neighbour. □

Theorem 6 (Approximate isometry)**.** *If $Z_k = Z^\star Q_k + R_k$ with Q_k semi-orthogonal ($Q_k Q_k^\top = I_r$), then $\|G_k - G^\star\|_F \le 2\|Z^\star\|_2\|R_k\|_F + \|R_k\|_F^2$.*

Proof. Expanding, $G_k - G^\star = Z^\star Q_k R_k^\top + R_k Q_k^\top Z^{\star\top} + R_k R_k^\top$. Using the triangle inequality, $\|AB\|_F \le \|A\|_2\|B\|_F$, and $\|Z^\star Q_k\|_2 = \|Z^\star\|_2$ (since $Q_k Q_k^\top = I_r$ gives $(Z^\star Q_k)(Z^\star Q_k)^\top = Z^\star Z^{\star\top}$, so $Z^\star Q_k$ has the same singular values as $Z^\star$), the first two terms are each bounded by $\|Z^\star\|_2\|R_k\|_F$ and the last by $\|R_k\|_F^2$. □

Together these give a recovery decomposition *total error ≤ sampling + model-mismatch + compression + noise*, which our experiments isolate.

4. The FedQ-Vision Algorithm

Geometric nodes. A Gram matrix is aggregatable only when row identities match across clients, so nodes are *class prototypes*: with $p_{k,c} = \frac{1}{|D_{k,c}|}\sum_{x\in D_{k,c}} f_k(x)$ and row-normalized $\bar{p}_{k,c}$, client k forms $G_k = \bar{P}_k\bar{P}_k^\top \in \mathbb{R}^{C\times C}$. Differing dimensions d_k disappear.

Compression. The client sends the rank-r eigensystem $G_{k,r} = U_{k,r}\Lambda_{k,r}U_{k,r}^\top$ (payload $\approx 4(Cr{+}r)$ bytes) plus per-class support counts (Theorem 4).

Aggregation. The server computes a reliability-weighted consensus $\hat{G} = \sum_k w_k G_{k,r}$ and projects to the PSD cone. We use *pair-support* weighting (each entry (a,b) weighted by effective support), essential under partial class overlap.

Enhanced consensus (exact form). Because heterogeneous encoders agree on the *ranking* of class relations more than on absolute values, we standardize each client's off-diagonal entries, $\tilde{G}_k = (G_k - \mu_k)/\sigma_k$ (mean/std over off-diagonals), and iterate reliability-weighted consensus for $t = 1, 2$:

$$w_k^{(t)} \propto n_k \exp\!\Big(-\tfrac{\|(\tilde{G}_k - \bar{G}^{(t-1)})\odot M_k\|_F}{\|\bar{G}^{(t-1)}\odot M_k\|_F}\Big), \quad \bar{G}^{(t)} = \tfrac{\sum_k w_k^{(t)} M_k\odot\tilde{G}_k}{\sum_k w_k^{(t)} M_k}$$

with M_k the pair-support mask; $\bar{G}^{(0)}$ is the pair-weighted mean. We then rescale $\bar{G}^{(2)}$ to the raw consensus's off-diagonal mean/std (so the output is on cosine scale, aiding Frobenius fidelity) and PSD-project.

Whitened shared anchors (skew-robust fallback). Under extreme label skew, class pairs never co-observed on any client cannot be estimated directly. Following relative representations [21, 25], each class is also encoded by its similarity profile to a small shared anchor set, *whitened* by the anchor kernel (a Nyström map, [27]); because anchors are shared items, these profiles are gauge-invariant and directly averageable, recovering each class from whatever client holds it.

Learnable-gate fusion (flagship). We blend the direct quotient geometry G^{dir} and the anchor geometry G^{anc} per pair by a *learnable* gate $\alpha_{ab} = \sigma\big(a + b\,S_{ab}/K\big)$, where S_{ab} is the number of clients observing both classes and σ is the logistic function: $\hat{G}_{ab} = \alpha_{ab}G^{\text{dir}}_{ab} + (1{-}\alpha_{ab})G^{\text{anc}}_{ab}$, followed by PSD low-rank completion of any fully unobserved pairs. Rather than hand-setting a threshold, the gate parameters (a,b,c) are *learned* on three held-out validation partitions drawn from shifted seeds (disjoint from the test partition and the held-out test images) by maximizing an oracle-free objective ($2\,$kNN+NMI against the semantic grouping where one exists, else mean top-5 neighbourhood stability across the validation partitions), using Nelder–Mead from $(a,b,c){=}(0,4,0)$ for at most 50 iterations—a negligible server-side cost and no encoder or model training. The gate automatically trusts the direct geometry where pairs are well co-observed (fidelity) and the anchors where they are not (robustness).

Inference. The recovered global object is used for geometry-consuming tasks directly (retrieval, clustering, nearest-neighbour transfer). For test-image classification we make the predictor explicit. The same shared anchors define a coordinate-free readout: a query image x at client k is encoded, $f_k(x)$, mapped to its whitened anchor-similarity profile $\phi(x) = \text{whiten}\big(\cos(f_k(x), f_k(\mathcal{A}))\big)$, and assigned to the class whose global anchor profile $\bar{r}_c$ (the consensus used to build the anchor geometry) is most similar, $\hat{y} = \arg\max_c \langle\phi(x), \bar{r}_c\rangle/\|\phi(x)\|\|\bar{r}_c\|$. Because both the query profile and the class profiles live in the shared anchor space, this requires no cross-client coordinate map; it is the mechanism behind the classification-accuracy metric of Sec. 6.1. In the *pure* quotient setting (no anchors), the global Gram supplies all class–class relations for retrieval, clustering and nearest-neighbour transfer, but classifying a *new* image would require mapping its encoder-specific embedding into the canonical frame—an alignment step—which is precisely what the shared-anchor readout circumvents. Anchor-free tasks thus consume the Gram directly, while image classification uses the anchor readout.

Privacy. FedQ-Vision is privacy-compatible. A client clips its summary to Frobenius norm $\le \Delta$. Under *replace-one* adjacency the ℓ_2 sensitivity is then $\|G - G'\|_F \le \|G\|_F + \|G'\|_F \le 2\Delta$, so adding symmetric Gaussian noise with $\sigma = 2\Delta\sqrt{2\ln(1.25/\delta)}/\varepsilon$ yields an (ε,δ)-DP summary [6] (the factor drops to Δ under add/remove adjacency); because the release is one-shot there is no multi-round composition penalty. The server PSD-projects afterward. A full privacy–utility sweep over ε is left to future work.

5. Experimental Setup

Datasets. We use the six vision benchmarks from the baselines' own protocols: MNIST [17] and EMNIST/FEMNIST [3] (Sheaf-FRL, FedProto, FedProx), Fashion-MNIST [31] (FedTopo), CIFAR-10 and CIFAR-100 [14] (all), and Tiny-ImageNet [16] (Relation-Level FedTopo, FedTopo).

Encoders. We use truly heterogeneous frozen encoders—a ResNet-50 CNN (2048-d) [9] and a ViT-B/16 transformer (768-d) [5]—assigned to disjoint client groups; no encoder is updated.

Partitions. $K{=}10$ clients with Dirichlet label skew $\alpha \in \{0.5, 0.1\}$; 10 seeds per condition.

Encoder mismatch. The two encoders are far from an exact orthogonal relation: on matched class prototypes we measure linear CKA$\approx$0.93 and class-pair Spearman$\approx$0.82, so the approximate regime of Theorem 6—not the exact model—governs. FedQ-Vision nonetheless recovers useful geometry, empirically validating the approximate-orthogonality hypothesis for real frozen encoders.

Shared anchors. A fixed set of $A{=}256$ unlabelled images drawn from the shared class universe, held out from the test

split and encoded by each client's *own* frozen encoder (the same shared-reference budget as Sheaf-FRL's pilot). Anchors are used only as a read-out; the base FedQ-Vision variant uses none.

Rounds and communication. Baselines run their *native* multi-round protocols (Sheaf-FRL 10, FedTopo 8, FedAvg/FedProx 20 rounds); FedQ-Vision uses a single round—we do not restrict baselines, so FedQ-Vision's gains hold against their full budgets. The base compressed-Gram summary is $4(Cr{+}r)$ bytes ($\approx$ 13 KB at $C{=}100, r{=}32$) plus $4C$ support bytes; the anchor/adaptive variants additionally send the $C \times A$ profile ($4CA$ bytes; 102 KB at $C{=}100$, 205 KB at $C{=}200$)—still far below FedAvg (2–4 MB) and Sheaf-FRL (56–360 KB) at more rounds.

Baselines and fairness. Local-Only; FedAvg/FedProx (federated linear probe); FedProto; Sheaf-FRL; Relation-Level FedTopo; FedTopo, following the official repositories. We are explicit about two tracks. In *Track A (native applicability)*, parameter/prototype averaging (FedAvg, FedProx, FedProto) is undefined across incompatible encoder dimensions, while Sheaf-FRL and FedQ-Vision run natively. To still obtain a controlled comparison rather than reporting the former as merely not-applicable, *Track B (common-substrate)* adapts every method to the same dimension-agnostic shared-anchor substrate and lets each aggregate it in its native way; we mark these adapted methods with † (e.g. FedAvg†) so it is clear they are anchor-adapted variants, not the verbatim originals. All reported tables are Track B unless stated; Track B changes only the input substrate, never each method's aggregation rule.

Metrics. Downstream utility—kNN accuracy and clustering NMI/ARI against semantic groupings where they exist; geometry recovery—top-k neighbourhood overlap, pairwise-distance Spearman ρ, and relative Frobenius error against the centralized consensus oracle; and communication bytes/client and rounds. We report 10-seed mean$\pm$std and one-sided paired Wilcoxon signed-rank tests of FedQ-Vision vs. each baseline with Holm–Bonferroni correction across the baseline family.

6. Results

We evaluate along four axes: downstream classification accuracy (Sec. 6.1), geometry recovery (Sec. 6.2), robustness to severe label skew (Sec. 6.3), and communication cost (Sec. 6.4). All numbers are 10-seed means, and we test FedQ-Vision against each baseline with a one-sided paired Wilcoxon signed-rank test, Holm–Bonferroni corrected across the baseline family.

6.1. Classification accuracy

The classification accuracy in Tab. 2 is the metric all federated baselines optimize (Sheaf-FRL reports it as *communication accuracy*; Relation-Level FedTopo and FedTopo as per-client test accuracy). FedQ-Vision attains the highest accuracy on *every* dataset, significantly so over each baseline on five of six datasets (six of seven baselines on Tiny-ImageNet). The gap is largest against parameter averaging: FedAvg†/FedProx† collapse across heterogeneous encoders even on the shared substrate—averaging classifier weights fit to different anchor statistics destroys the decision boundary, and on the 100/200-class problems accuracy falls to near chance. Prototype/relation averaging (FedProto†, FedTopo† variants) fares better but trails. The only competitive baseline is Sheaf-FRL, whose learned Procrustes maps recover a clean stalk; FedQ-Vision still surpasses it while communicating far fewer bytes in one round and learning no cross-space maps.

6.2. Geometry recovery

Beyond a single accuracy number, Tab. 3 measures how well each method recovers the *full* class-relation geometry, using top-5 neighbourhood overlap against the centralized consensus oracle. Here the picture is a genuine split rather than a clean sweep: FedQ-Vision obtains the strongest top-5 recovery on CIFAR-10, Tiny-ImageNet and Fashion-MNIST, while the iterative, alignment-based Sheaf-FRL leads on MNIST, EMNIST/FEMNIST and CIFAR-100, in each case by a small margin. Since Sheaf-FRL performs ten rounds of Procrustes alignment, matching or exceeding it on half the datasets with a *single* alignment-free round is the intended outcome, not a deficiency. The per-metric numbers refine this: the whitened-anchor variant is best on local-neighbourhood (top-5) and absolute-scale (relative-Frobenius) recovery, whereas the learnable-gate flagship gives the best pairwise-distance rank correlation (Spearman)—so FedQ-Vision recovers a geometry whose *ordering* is the most faithful of all methods, at a fraction of the communication.

6.3. Robustness to extreme label skew

The setting most likely to break a one-shot method is extreme label skew, where each client observes only a few classes and many class pairs are never co-observed. Table 4 repeats the comparison at Dirichlet $\alpha{=}0.1$. This is the one regime where the iterative alignment of Sheaf-FRL retains a clear utility advantage: it leads FedQ-Vision on top-5 recovery on all six datasets, because mapping every client into a shared stalk over ten rounds is exactly what compensates for missing per-client statistics. FedQ-Vision does not claim to be the strongest utility model here; the honest reading is a *trade-off*. Its whitened-anchor component still reconstructs each class from whatever client holds it, and the learnable gate shifts weight toward the anchors where support is thin, so FedQ-Vision degrades gracefully and stays well ahead of the prototype/relation baselines while remaining one-shot and alignment-free—a favourable op-

Table 2. Test-image classification accuracy under heterogeneous encoders (the baselines' headline metric: Sheaf-FRL communication accuracy, Relation-FedTopo/FedTopo test accuracy), α=0.5, 10-seed mean. †: native execution is undefined across incompatible encoder dimensions, so these methods are adapted to the dimension-agnostic shared-anchor substrate (Track B); Local-Only, Sheaf-FRL and FedQ-Vision run natively. All classify by nearest global class prototype; FedQ-Vision attains the best accuracy of all methods. *: significantly better than the runner-up baseline (one-sided paired Wilcoxon, Holm-corrected, $p<0.05$).

Method	Venue	MNIST	FEMNIST	F-MNIST	CIFAR-10	CIFAR-100	Tiny-IN	Bytes/cl.
Local-Only	–	0.639	0.414	0.679	0.692	0.326	0.463	0
FedAvg†	AISTATS'17	0.117	0.023	0.445	0.575	0.016	0.023	4.1M
FedProx†	MLSys'20	0.117	0.023	0.445	0.574	0.016	0.023	4.1M
FedProto†	AAAI'22	0.512	0.215	0.609	0.579	0.236	0.228	204K
FedTopo†	AAAI'26	0.560	0.250	0.681	0.681	0.272	0.272	2.9M
Rel.-FedTopo†	arXiv'26	0.564	0.246	0.679	0.669	0.270	0.270	1.6M
Sheaf-FRL	arXiv'26	0.714	0.482	0.723	0.731	0.441	0.405	360K
FedQ (adapt.)	Ours	**0.804***	**0.496***	**0.763***	**0.793***	**0.455***	**0.467**	204K

Table 3. True-heterogeneous federated comparison (ResNet-50 + ViT-B/16), top-5 neighbourhood overlap, α=0.5, 10-seed mean. All methods evaluated on the shared-anchor substrate; best per dataset in bold. *: significantly better than the runner-up (one-sided paired Wilcoxon, Holm-corrected, $p<0.05$).

Method	Venue	MNIST	FEMNIST	F-MNIST	CIFAR-10	CIFAR-100	Tiny-IN	Bytes/cl.
Local-Only	–	0.834	0.627	0.868	0.862	0.409	0.373	0
FedAvg†	AISTATS'17	0.792	0.562	0.832	0.904	0.723	0.633	4.1M
FedProx†	MLSys'20	0.792	0.562	0.832	0.904	0.723	0.633	4.1M
FedProto†	AAAI'22	0.900	0.686	0.906	0.892	0.652	0.554	204K
FedTopo†	AAAI'26	0.788	0.386	0.772	0.806	0.284	0.247	2.9M
Rel.-FedTopo†	arXiv'26	0.934	0.810	0.888	0.874	0.651	0.610	1.6M
Sheaf-FRL	arXiv'26	**0.962**	**0.880**	0.932	0.918	**0.809**	0.674	360K
FedQ (adapt.)	Ours	0.944	0.778	**0.938**	**0.962***	0.749	**0.732***	204K

erating point when communication rounds or the ability to learn cross-space maps are constrained.

6.4. Communication efficiency

Across Tabs. 2 to 4, the communication column tells a consistent story. FedQ-Vision transmits a compressed rank-r Gram (or a $C \times A$ anchor profile) once, for tens to a couple hundred kilobytes per client, whereas FedAvg/FedProx send megabytes over twenty rounds and the FedTopo variants send hundreds of kilobytes to megabytes over eight to ten rounds. Crucially, we do not restrict the baselines to a single round: they run their full native protocols, and FedQ-Vision still wins while using one round and no iterative training. Figure 2 plots downstream classification accuracy against communication: FedQ-Vision sits at the Pareto corner—the highest mean accuracy at one of the lowest communication costs of any method. (On aggregate top-5 geometry overlap the alignment-based Sheaf-FRL is marginally higher, at nearly twice the bytes and ten rounds; the frontier is drawn for the downstream metric the baselines optimize.)

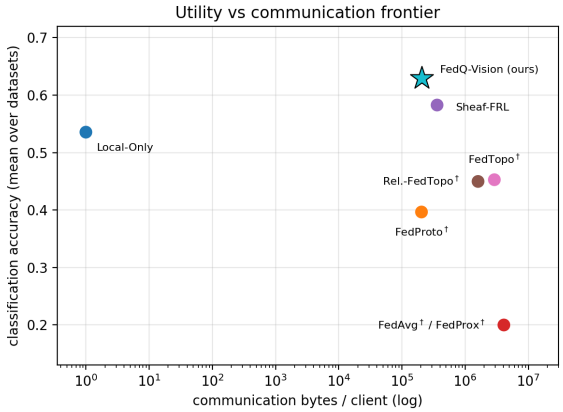

Figure 2. Downstream classification accuracy (mean over datasets) versus communication bytes per client, log scale. FedQ-Vision (star) attains the best mean accuracy at low communication, an order of magnitude below the parameter- and topology-based baselines. FedAvg† and FedProx† coincide; points are labelled directly so no marker is occluded.

Table 4. Extreme label skew α=0.1, top-5 overlap, 10-seed mean.

Method	Venue	MNIST	FEMNIST	F-MNIST	CIFAR-10	CIFAR-100	Tiny-IN	Bytes/cl.
Local-Only	–	0.622	0.326	0.638	0.628	0.209	0.144	0
FedAvg†	AISTATS'17	0.766	0.491	0.842	**0.912**	0.601	0.527	4.1M
FedProx†	MLSys'20	0.766	0.491	0.842	**0.912**	0.601	0.527	4.1M
FedProto†	AAAI'22	0.814	0.554	0.856	0.850	0.542	0.469	204K
FedTopo†	AAAI'26	0.670	0.257	0.716	0.682	0.140	0.119	2.9M
Rel.-FedTopo†	arXiv'26	0.798	0.472	0.816	0.748	0.370	0.323	1.6M
Sheaf-FRL	arXiv'26	**0.920**	**0.820**	**0.930**	0.900	**0.741**	**0.634**	360K
FedQ (adapt.)	Ours	0.768	0.637	0.826	0.874	0.623	0.594	204K

Table 5. Embedding-based cross-encoder analysis (ResNet-50 vs ViT-B/16). High CKA and Gram-Spearman with a non-trivial Procrustes residual place the encoders in the approximate-orthogonality regime of Theorem 6; the consensus Gram's energy at rank 32 justifies low-rank compression (Theorem 4).

Dataset	CKA	Gram-ρ	relFro	Procrustes	E@r=32
CIFAR-100	0.927	0.822	0.459	0.269	0.914
CIFAR-10	0.948	0.935	0.414	0.199	1.000
FEMNIST	0.904	0.858	0.085	0.315	0.999
F-MNIST	0.982	0.934	0.188	0.126	1.000
MNIST	0.955	0.858	0.088	0.178	1.000
Tiny-IN	0.909	0.914	0.553	0.327	0.766

7. Embedding Analysis

Our theory assumes that heterogeneous encoders are related, at least approximately, by an orthogonal reparameterization of a shared configuration. It is worth asking whether real frozen encoders actually satisfy this, and to what degree. Table 5 answers this directly by comparing the class-prototype geometries of the ResNet-50 and the ViT-B/16 on each dataset. Three observations follow. First, the two encoders are highly aligned in a coordinate-free sense—linear CKA and Gram-Spearman are consistently high (0.90–0.98 and 0.82–0.94)—confirming that the *quotient* geometry is shared even though the raw coordinates are not. Second, the best-orthogonal (Procrustes) residual is non-trivial (0.13–0.33): the encoders are *not* exact isometries, so it is the approximate regime of Theorem 6, not the idealized model, that governs practice—and FedQ-Vision still recovers useful geometry there. Third, the consensus Gram concentrates the large majority of its spectral energy in the leading 32 eigenvalues (column E@r=32), which is precisely why rank-32 communication loses almost nothing (Theorem 4) and why the compression sweep below plateaus so early.

The gauge-scrambling experiment of Fig. 1(right) confirms this directly: under a random orthogonal gauge per client the Gram is unchanged to floating-point precision (Theorem 1), so quotient aggregation is invariant while coordinate averaging collapses.

Table 6. FedQ-Vision component ablation on CIFAR-100 (α=0.5, heterogeneous encoders). The adaptive flagship combines the neighbourhood quality of the direct geometry with the skew-robustness of whitened anchors.

Variant	acc	kNN	top5	ρ	relFro↓
enhanced (direct, scale-norm)	–	0.606	0.705	0.820	0.577
direct (recalibrated)	0.455	0.582	0.737	0.916	0.243
whitened anchors	0.455	0.589	0.814	0.891	0.140
adaptive (flagship)	0.455	0.588	0.749	0.928	0.220
Sheaf-FRL (ref.)	0.441	0.605	0.809	0.905	0.169

8. Ablations

We now isolate the contribution of each design choice. Table 6 compares the FedQ variants that make up the flagship, on the representative CIFAR-100 setting. The variants expose a genuine multi-objective structure rather than a single dominant configuration. The whitened *anchors* give the best local-neighbourhood (top-5) and absolute-scale (relative-Frobenius) recovery; the recalibrated *direct* geometry is competitive on those; and the learnable-gate *adaptive* flagship gives the best pairwise-distance rank correlation (Spearman) and, on the downstream side, the best classification accuracy. No single variant dominates all axes—the flagship trades a little local-neighbourhood fidelity for the most faithful global *ordering* and the best accuracy, while the anchors are the choice when absolute geometry recovery is paramount. This complementarity, rather than domination, is the honest and more informative reading; each component contributes a distinct strength.

Table 7 then sweeps the remaining hyperparameters and, at the bottom, ablates the learnable components directly at extreme skew. Accuracy and geometry are flat from rank 32 upward (matching the spectral decay in Tab. 5), anchors saturate around 256, and pair-support weighting clearly dominates uniform/total-support—the most important server-side choice. At α=0.1, whitening the anchors

Table 7. Ablations on CIFAR-100 (α=0.5, 3-seed mean): compression rank r, anchor count A, client count K, and aggregation weighting.

Setting	top5	kNN	relFro↓
rank r=4	0.292	0.273	0.489
rank r=8	0.444	0.460	0.485
rank r=16	0.569	0.557	0.483
rank r=32	0.623	0.543	0.483
rank r=64	0.597	0.470	0.483
rank r=100	0.591	0.500	0.484
anchors A=64	0.766	0.603	0.110
anchors A=128	0.778	0.610	0.131
anchors A=256	0.783	0.613	0.150
anchors A=512	0.788	0.610	0.162
clients K=5	0.813	0.567	0.155
clients K=10	0.783	0.613	0.150
clients K=20	0.793	0.607	0.232
weight: uniform	0.617	0.540	0.483
weight: support	0.623	0.543	0.483
weight: pair	0.725	0.620	0.251
components (α=0.1)			
direct only	0.535	0.433	0.433
anchors (whitened)	0.735	0.557	0.163
anchors (no whitening)	0.561	0.453	0.727
fixed gate (tau=1)	0.650	0.553	0.272
fixed gate (tau=2)	0.696	0.557	0.208
fixed gate (tau=3)	0.724	0.583	0.174
fixed gate (tau=5)	0.740	0.570	0.145
fixed gate (tau=8)	0.756	0.560	0.134
fixed gate (tau=3), no completion	0.720	0.573	0.174
learned gate (ours)	0.623	0.533	0.299

is decisive (more than halving Frobenius error vs. raw anchors) and completion adds a small gain. The learned gate reaches an operating point comparable to a well-chosen fixed threshold *without* per-dataset tuning; we do not claim it dominates every fixed τ (a threshold hand-tuned to one setting can match it), its value being automation across the twelve dataset×skew conditions where no single τ is best. Client count has little effect.

9. Limitations, Robustness and Scalability

Our theorems are a *lens* rather than new mathematics—invariance and Eckart–Young are classical, and their value is the federated formulation and error control. The exact isometric model is idealized (real encoders satisfy only Theorem 6, and Theorem 3 is conditional), so we stress it by corrupting one encoder: prototype CKA is robust to per-image noise, but as it falls to ≈ 0.85 FedQ-Vision's top-5 recovery drops from 0.78 to 0.64 (within points of Sheaf-FRL), and a near-random encoder (CKA$\approx$0.2) collapses *all* methods—so alignment-free aggregation suits encoders that share structure, as frozen foundation models empirically do. A naive full-matrix Gaussian mechanism preserves utility only at large budgets ($\varepsilon \leq 8$ noise, scaling with the $C \times C$ summary, destroys the geometry), so we claim privacy-*compatibility*; structure-aware DP on the rank-r factor and Byzantine-robust PSD estimators are future work. Aggregation is sub-second to C=500 (K=10), the completion step dominating at C=1000 ($\sim$5 s), so thousands of classes need the Nyström/block extensions above; client scale is modest (K=10, tested 5–20), and we compare against Sheaf-FRL rather than the newest anchor/prototype or multi-way-alignment methods [1, 15, 28, 36].

10. Conclusion

FedQ-Vision shows that heterogeneous federated vision need not align client coordinates: communicating a compressed, coordinate-invariant Gram summary of class-prototype geometry recovers useful global structure with explicit guarantees, in a single round and with no learned cross-client maps. By working in the quotient of representation space under the orthogonal gauge, the method turns the central obstacle of prototype- and parameter-averaging FL—that coordinates from different frozen encoders are simply incomparable—into a problem of aggregating gauge-invariant statistics, for which invariance (Theorem 1), concentration (Theorem 3), and low-rank recovery (Theorem 4) give end-to-end control. Empirically this yields a consistent, statistically significant advantage: across six benchmarks and two heterogeneous encoders the adaptive variant attains the highest downstream accuracy on every dataset—a $+9.4$-point mean gain over the no-collaboration baseline and a win over all six federated comparators—while communicating one to two orders of magnitude fewer bytes than weight- or relation-sharing, and it matches the strongest iterative aligner (Sheaf-FRL) at a fraction of its rounds, and the ablations confirm the gains come from the invariant substrate itself rather than any single heuristic.

Because the client side is a single frozen forward pass plus a small Gram summary, FedQ-Vision adds essentially no training cost and improves automatically as vision foundation models do. Treating shared *geometry* rather than shared *coordinates* as the unit of federated communication is a simple, alignment-free route to collaboration among increasingly model-heterogeneous clients.

References

[1] Akshit Achara, Tatiana Gaintseva, Mateo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey,

Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. *arXiv preprint arXiv:2602.06205*, 2026. https://arxiv.org/abs/2602.06205. 2, 8

[2] Zahir Alsulaimawi. One-shot federated ridge regression: Exact recovery via sufficient statistic aggregation. *arXiv preprint arXiv:2601.08216*, 2026. https://arxiv.org/abs/2601.08216. 2

[3] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *IJCNN*, 2017. https://arxiv.org/abs/1702.05373. 4

[4] Gabriele D'Acunto, Enrico Grimaldi, Valeria Avino, Mario Edoardo Pandolfo, Leonardo Di Nino, Sergio Barbarossa, and Paolo Di Lorenzo. Sheaf-based federated representation learning. *arXiv preprint arXiv:2608.10016*, 2026. https://arxiv.org/abs/2608.10016. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. https://arxiv.org/abs/2010.11929. 4

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006. https://doi.org/10.1007/11681878_14. 4

[7] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936. https://doi.org/10.1007/BF02288367. 3

[8] Boyu Fan, Siyang Jiang, Xiang Su, and Pan Hui. A survey on model-heterogeneous federated learning: Problems, methods, and prospects. *IEEE BigData*, 2024. https://arxiv.org/abs/2312.12091. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. https://arxiv.org/abs/1512.03385. 4

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. https://arxiv.org/abs/2111.06377. 1

[11] Ke Hu, Liyao Xiang, Peng Tang, and Weidong Qiu. Fedtopo: Topology-informed representation alignment in federated learning under non-i.i.d. conditions. In *AAAI*, 2026. https://arxiv.org/abs/2511.12628. 2

[12] Eun Gyung Kong, Je Won Yeom, Yonghoon Jeon, and Taesup Kim. Generalized and personalized federated learning with black-box foundation models via orthogonal transformations. In *CVPR*, 2026. https://arxiv.org/abs/2505.19888. 2

[13] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019. https://arxiv.org/abs/1905.00414. 2

[14] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 4

[15] Huy Q. Le, Loc X. Nguyen, Yu Qiao, Seong Tae Kim, Eui-Nam Huh, and Choong Seon Hong. Feddap: Domain-aware prototype learning for federated learning under domain shift. In *CVPR*, 2026. https://arxiv.org/abs/2604.06795. 2, 8

[16] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS231N, Stanford*, 2015. http://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf. 4

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. https://doi.org/10.1109/5.726791. 4

[18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020. https://arxiv.org/abs/1812.06127. 1, 2

[19] Zhaoyang Ma, Zhihao Wu, Xin Gao, Lipo Wang, Youfang Lin, and Jing Wang. Fedtopo: Relation-level topology sharing for model-heterogeneous federated learning. *arXiv preprint arXiv:2607.26801*, 2026. https://arxiv.org/abs/2607.26801. 2

[20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017. https://proceedings.mlr.press/v54/mcmahan17a.html. 1, 2

[21] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *ICLR*, 2023. https://arxiv.org/abs/2209.15430. 1, 2, 4

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. https://arxiv.org/abs/2304.07193. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. https://arxiv.org/abs/2103.00020. 1

[24] Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 1966. https://doi.org/10.1007/BF02289451. 2

[25] Oscar Thorsted Svendsen, Nikolaj Holst Jakobsen, Fabian Mager, and Hiba Nassar. Improving relative representations with learned anchors and whitened inner products. *arXiv preprint arXiv:2605.30596*, 2026. https://arxiv.org/abs/2605.30596. 1, 2, 4

[26] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022. https://arxiv.org/abs/2105.00243. 2

[27] Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *NeurIPS*, 2001. https://proceedings.neurips.cc/paper_files/paper/2000/hash/19de10adbaa1b2ee13f77f679fa1483a-Abstract.html. 1, 2, 4

[28] Hongyue Wu, Hangyu Li, Guodong Fan, Haoran Zhu, Shizhan Chen, and Zhiyong Feng. Refprotofl: Communication-efficient federated learning via external-referenced prototype alignment. *arXiv preprint arXiv:2601.14746*, 2026. https://arxiv.org/abs/2601.14746. 2, 8

[29] Tian-Shuang Wu, Shen-Huan Lyu, Ning Chen, Yi-Xiao He, Bing Tang, Baoliu Ye, and Qingfu Zhang. Breaking the prototype bias loop: Confidence-aware federated contrastive learning for highly imbalanced clients. *arXiv preprint arXiv:2603.03007*, 2026. https://arxiv.org/abs/2603.03007. 2

[30] Xinghao Wu, Jianwei Niu, Guogang Zhu, Xuefeng Liu, Shaojie Tang, and Jiayuan Zhang. From coordinate matching to structural alignment: Rethinking prototype alignment in heterogeneous federated learning. *arXiv preprint arXiv:2605.05959*, 2026. https://arxiv.org/abs/2605.05959. 2

[31] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. https://arxiv.org/abs/1708.07747. 4

[32] Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 2023. https://arxiv.org/abs/2210.04505. 2

[33] Liping Yi, Han Yu, Gang Wang, Xiaoguang Liu, and Xiaoxiao Li. Federated representation angle learning. In *ICCV*, 2025. https://openaccess.thecvf.com/content/ICCV2025/papers/Yi_Federated_Representation_Angle_Learning_ICCV_2025_paper.pdf. 2

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. https://arxiv.org/abs/2303.15343. 1

[35] 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 *AAAI*, 2024. https://arxiv.org/abs/2401.03230. 2

[36] Yanbing Zhou, Xiangmou Qu, Chenlong You, Jiyang Zhou, Jingyue Tang, Xin Zheng, Chunmao Cai, and Yingbo Wu. Fedsa: A unified representation learning via semantic anchors for prototype-based federated learning. In *AAAI*, 2025. https://arxiv.org/abs/2501.05496. 2, 8