Class Prototype Alignment Concentrates Chest X-ray Representations in Federated and Pooled Training

Jingjing Yang[1⋆] and Xia Zhao[1]

School of Information Science and Engineering, Hebei North University, Zhangjiakou 075000, China
landryyang@hebeinu.edu.cn

Abstract. Prototype-based federated learning represents each class with compact feature summaries, but alignment to a single mean may suppress within-class variation in heterogeneous images. We test this mechanism in federated ChestX-ray14 classification and ask whether federation is necessary. Across three seeds, two patient-level client allocations, and weight-averaging intervals $K \in \{2, 5, 20\}$, we compare class-mean alignment with matched averaging-only controls. Alignment lowers the mean representation effective rank by 84–157 dimensions and the mean covariance effective rank by 114–129 dimensions across the six federated designs. Predictive effects vary across settings. Mean paired differences range from -0.0165 to $+0.0073$ in area under the receiver operating characteristic curve, or AUROC, and from -0.0022 to $+0.0114$ in average precision. Three pooled training pairs reproduce the concentration. Their mean log ratio of aligned to control covariance effective rank is -4.4241, and their mean AUROC difference is -0.0444. On 3,000 VinDr-CXR images, all 21 paired comparisons show lower effective ranks and a larger leading variance fraction after alignment. Class-mean alignment concentrates representations across training designs, but its geometry does not consistently track predictive benefit.

Keywords: Federated learning · Chest X-ray classification · Prototype alignment · Effective rank · Representation geometry.

1 Introduction

Federated learning (FL) trains models across distributed data without centralizing raw records [9,6]. In medical imaging, data often remain within clinical institutions [16]. Standard FL communicates model updates, while prototype-based methods exchange compact class summaries that can support clients with different model architectures [18,15].

A single class prototype is efficient, but its geometry is restrictive. A quadratic alignment loss pulls every feature toward the same class mean even when a label

⋆ Corresponding author.

contains substantial anatomical and acquisition variation. Recent methods alter the number, geometry, or timing of prototypes to address limitations of direct class-mean matching [15,2,21,4,20,1]. The effect of direct class-mean alignment on learned dimensionality remains poorly characterized in medical FL.

Representation spectra are widely used to study dimensional collapse [5]. Effective rank provides a continuous summary of spectral concentration [17]. Because preprocessing and spectral weighting affect its value, predictive performance must be measured separately.

We therefore ask whether class-mean alignment concentrates chest X-ray representations and whether any such effect requires federation. Matched federated experiments isolate alignment while holding model averaging, training code, and evaluation data constant. Pooled controls then remove federation from the design. A label-free analysis on VinDr-CXR tests whether the geometry persists on images from an independent source. Centered spectra and predictive ranking are evaluated separately throughout.

2 Related Work

Prototype-based FL communicates class-level representations rather than relying only on model parameters. Federated averaging shares model parameters, whereas FedProto regularizes local representations toward aggregated class means [9,18]. Recent methods replace the single prototype, constrain prototype geometry, or change when alignment begins. MP-FedCL retains multiple prototypes per class, while FedNH combines class semantics with uniformly separated prototypes [15,2]. FedSA and FedLSA learn semantic anchors. FedLSA also uses hyperspherical contrast for domain-skewed data [21,4]. Recent preprints align inter-class relations or delay alignment until local representations mature [20,1]. Our question concerns the geometry induced by direct class-mean alignment in a matched medical imaging experiment.

Dimensional collapse denotes the loss of variation along feature directions [5]. Roy and Vetterli define effective rank from normalized singular values [17]. We report matrix and covariance forms after centering because singular values and their squares weight the spectrum differently. Neural collapse is a broader terminal-training phenomenon that combines vanishing within-class variability, simplex equiangular class means, and classifier alignment [14]. Neural-collapse-inspired FL has used fixed simplex equiangular tight-frame classifiers to regularize federated representations [7]. The present analysis concerns held-out representation spectra and within-class variation rather than the full neural-collapse geometry.

3 Methods

3.1 Task, cohorts, and client allocation

ChestX-ray14 [19,13] is reduced to Pneumonia versus No Finding. Images with neither label are excluded. Positive images may carry additional findings, and the

Table 1. Study cohorts. Positive denotes Pneumonia and negative denotes No Finding.

Split	Images	Patients	Positive	Pos. rate	PA / AP
Train	46,239	20,250	802	1.73%	31,407 / 14,832
Validation	5,137	2,226	74	1.44%	3,558 / 1,579
Test	10,416	2,576	555	5.33%	4,967 / 5,449

labels were extracted from radiology reports rather than assigned prospectively. We retain the official patient-disjoint test split and assign 10% of the official training patients to validation with a deterministic rule. No patient appears in more than one split. View position is reported as posteroanterior (PA) or anteroposterior (AP). Table 1 summarizes the resulting cohorts. Average precision accompanies AUROC because test prevalence differs from training prevalence.

We construct two patient-level client allocations. The Dirichlet allocation draws class-specific proportions with $\alpha = 1$ across four clients. Across seeds, clients contain 424 to 26,339 images and have Pneumonia prevalence between 1.11% and 4.75%. The view-based allocation assigns each patient to a PA-majority or AP-majority client. The PA-majority client contains 28,941 PA and 4,014 AP images. The AP-majority client contains 10,818 AP and 2,466 PA images. Both clients contain mixed views. Training uses class-balanced sampling within each client.

3.2 Training and matched comparisons

Every client starts from the same ImageNet-pretrained Vision Transformer model `vit_small_patch16_224` [3]. The encoder produces a 384-dimensional representation followed by a new binary head. Grayscale images are repeated across three channels and resized on the short edge to 255 pixels. Training uses a random 224×224 crop and horizontal flip. Evaluation uses a center crop, and all images receive ImageNet normalization.

All clients participate in each of 20 rounds and train for two local epochs per round. We use batches of 128 and AdamW with a learning rate of 5×10^{-5} and weight decay of 10^{-5} [8]. Gradients are clipped at 1.0. The optimizer is recreated for each local training call. Round 1 uses no prototype loss and establishes the first global prototypes. Each run uses one 24-GB NVIDIA RTX 3090 Ti. The primary federated models were trained with Python 3.12.3, PyTorch 2.5.1, and CUDA 12.1.

From round 2, the aligned condition minimizes binary cross-entropy (BCE) with a prototype-alignment term

$$L = L_{\mathrm{BCE}} + \lambda L_{\mathrm{align}}, \qquad L_{\mathrm{align}} = \frac{1}{Bd} \sum_{i=1}^{B} \|z_i - \hat{\mu}_{y_i}\|_2^2, \tag{1}$$

In Eq. 1, B is the minibatch size, z_i is the representation of sample i, and y_i is its label. The representation dimension is $d = 384$, and $\hat{\mu}_{y_i}$ is the corresponding

global prototype. We set $\lambda = 0.1$. At the end of each round, clients recompute unaugmented class means on their local training data. The server averages these means using class support. The matched control uses the same procedure with $\lambda = 0$.

Both conditions average client models by sample count after rounds divisible by $K \in \{2, 5, 20\}$. They perform 10, 4, and 1 model-averaging events. When $K = 20$, model averaging occurs only after the final local update.

3.3 Predictive evaluation

Every final checkpoint is evaluated on the same fixed test set. A threshold targeting 90% specificity is selected on validation for each client and transferred unchanged to test. We report sensitivity, specificity, positive predictive value, negative predictive value, Brier score, and ten-bin expected calibration error in the repository. The main analysis focuses on threshold-free AUROC and non-interpolated average precision, computed as precision weighted by the increase in recall across successive thresholds. Identical client states are verified by exact tensor comparison and evaluated once. Otherwise, client metrics are averaged within a run before aggregation across seeds.

We report all three seed values and summarize paired differences between aligned and averaging-only conditions. With three seeds, we use descriptive comparisons and do not conduct inferential tests. We repeat the comparison on the held-out validation set and on a sensitivity set that excludes the 5,615 images selected by prior test-set diagnostics.

3.4 Centered representation geometry

Let $\bar{z}$ be the row mean of the representation matrix Z, let $\widetilde{Z} = Z - \mathbf{1}\bar{z}^\top$, and let s_j denote the singular values of $\widetilde{Z}$. We compute

$$\begin{aligned} r_Z &= \exp\left(-\sum_j p_j \log p_j\right), \quad p_j = \frac{s_j}{\sum_\ell s_\ell}, \\ r_C &= \exp\left(-\sum_j q_j \log q_j\right), \quad q_j = \frac{s_j^2}{\sum_\ell s_\ell^2}. \end{aligned} \tag{2}$$

Equation 2 defines r_Z as the effective rank of the centered representation matrix and r_C as the effective rank of its covariance. We also measure the leading variance fraction q_1, class-conditional effective ranks, and within-class covariance trace. The ImageNet-pretrained encoder before federated training provides the initialization reference.

3.5 Pooled and external controls

The pooled and external analyses were specified before the additional models were trained. To separate alignment from federation, we trained pooled models

with a single learner. Each seed paired $\lambda = 0.1$ with $\lambda = 0$. The learner received the complete 46,239-image training cohort and otherwise followed the same architecture, augmentations, class-balanced sampling, 20 rounds, two epochs per round, optimizer reset, warm-up, and final-checkpoint evaluation. Because incomplete batches were dropped, pooled training used 92,416 sampled instances per round, 0.28–0.84% more than federated training. The pooled implementation used an initial mixed-precision loss scale of 1,024 and Python 3.11.15. PyTorch and CUDA versions matched the federated runs.

We evaluated label-free geometry on the 3,000-image VinDr-CXR test cohort [12,11]. The same deterministic transform was applied to every image. We measured r_Z, r_C, the leading variance fraction q_1, and covariance trace for all 36 federated and six pooled final checkpoints. We evaluated consistency by asking whether every seed-paired aligned model had lower r_Z and r_C and higher q_1 than its control. Because VinDr-CXR labels were not used, this analysis estimates geometry rather than external performance, prevalence, or class-conditional effects. All primary and extension checkpoints were evaluated in the same Python 3.11.15, PyTorch 2.5.1, and CUDA 12.1 environment. For each pair, $\Delta \log r_C = \log(r_C^{\text{aligned}}/r_C^{\text{control}})$, so negative values denote concentration.

4 Results

4.1 Predictive effects vary across metrics and settings

The direction of predictive change depends on both metric and experimental setting. Across the six matched settings, mean paired differences range from -0.0165 to $+0.0073$ in AUROC and from -0.0022 to $+0.0114$ in average precision. Table 2 gives the condition means, and Fig. 1 shows every paired seed. The largest metric divergence occurs under Dirichlet allocation at $K = 20$. AUROC differences are negative in all three seeds by 0.0009, 0.0144, and 0.0342, while average precision increases by 0.0030, 0.0064, and 0.0030.

Table 2. Prototype alignment has mixed predictive effects on the fixed test set. Entries are means across three seeds and Δ is prototype alignment minus averaging only.

Allocation	K	Alignment		Averaging only		ΔAUROC	ΔAvg. prec.
		AUROC	Avg. prec.	AUROC	Avg. prec.		
Dirichlet	2	.6916	.1216	.6982	.1178	−.0066	+.0038
Dirichlet	5	.6976	.1243	.7057	.1199	−.0081	+.0044
Dirichlet	20	.7175	.1303	.7340	.1262	−.0165	+.0041
View based	2	.6788	.1198	.6870	.1117	−.0082	+.0081
View based	5	.6851	.1189	.6778	.1075	+.0073	+.0114
View based	20	.7036	.1148	.7080	.1169	−.0044	−.0022

The validation comparison shows the same metric dependence. Across the six settings, mean differences range from -0.039 to $+0.023$ for AUROC and from

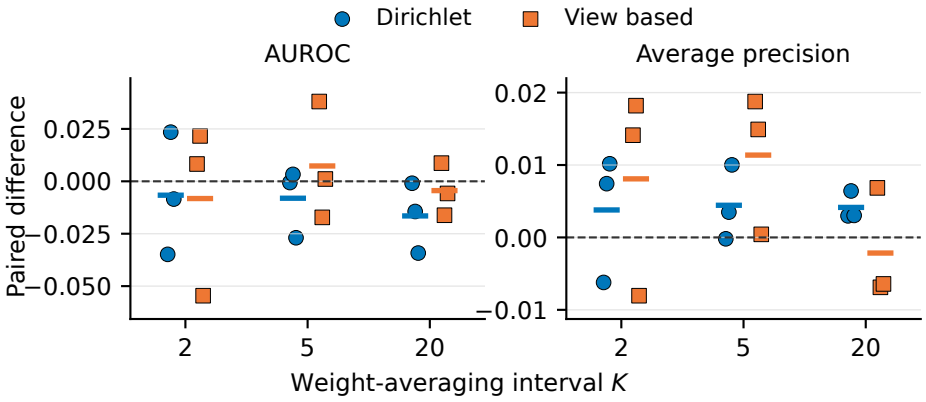

Fig. 1. Predictive effects of prototype alignment vary across metrics, client allocations, and seeds on the fixed test set. Positive values favor prototype alignment. Each point represents one seed, and each horizontal segment shows the three-seed mean.

-0.015 to $+0.010$ for average precision. On the image-exclusion sensitivity set, the corresponding ranges are -0.021 to $+0.020$ and -0.006 to $+0.009$.

4.2 Prototype alignment concentrates centered representations

Unlike the predictive metrics, both effective-rank measures shift in the same direction in every matched setting. At initialization, $r_Z = 259.5$ and $r_C = 72.7$. Averaging-only checkpoints have $r_Z = 299.9$–317.1 and $r_C = 106.6$–150.0. Aligned checkpoints have $r_Z = 123.7$–255.4 and $r_C = 1.5$–17.8. Their leading variance fraction is 0.51–0.95, compared with 0.09–0.17 for averaging alone. Alignment reduces the mean r_Z by 84–157 dimensions and the mean r_C by 114–129 dimensions in every matched allocation and K setting. Figure 2 shows the corresponding rank ratios to initialization.

Class-conditional analysis shows the same pattern. For negative images, aligned covariance effective ranks span 1.6–20.2, compared with 108.4–151.6 for averaging alone. For positive images, the corresponding ranges are 1.3–6.7 and 79.8–113.8. Across settings, mean within-class covariance trace ranges from 3.3 to 38.7 after alignment and from 939.8 to 1014.1 in the controls. All 18 paired differences are negative, showing that prototype alignment reduces within-class variance across both labels.

4.3 Concentration persists without federation and across image sources

Pooled training reproduces the geometric effect in every seed. Aligned models have $r_C = 1.42$–1.58, compared with 119.93–126.62 for their controls, and the mean $\Delta \log r_C$ is -4.4241. AUROC differences are negative by 0.0083–0.0952

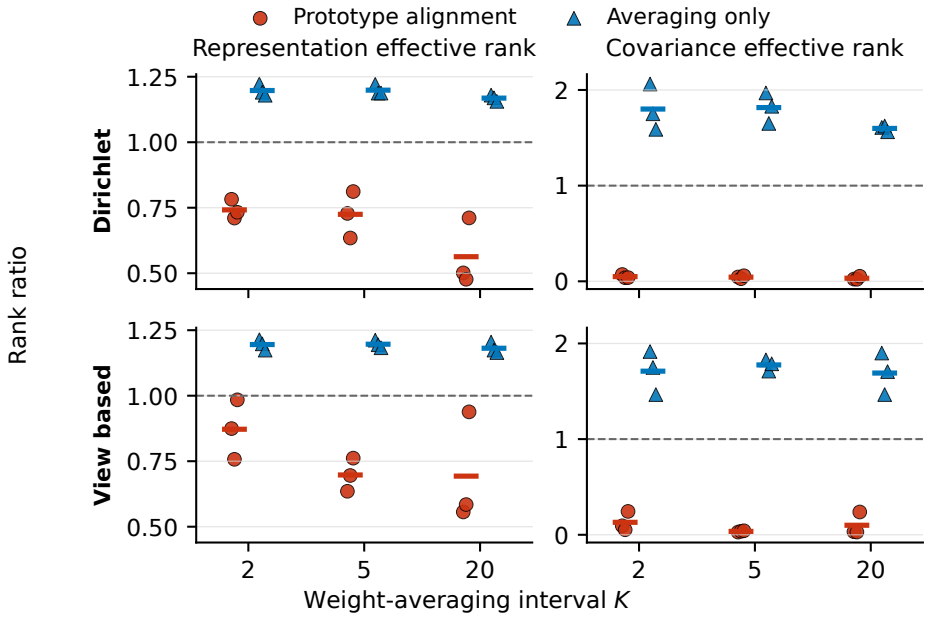

Fig. 2. Prototype alignment reduces both effective-rank measures across client allocations and averaging intervals on the fixed test set. Points represent seeds, horizontal segments show means, and the dashed line marks the ImageNet-pretrained initialization.

across the three pairs, with a mean difference of −0.0444. Average-precision differences range from −0.0180 to +0.0042 and average −0.0078. Concentration can therefore emerge without client partitioning or model aggregation.

Table 3. Alignment effects on covariance effective rank. Entries are mean $\Delta \log r_C$ with the seed range in brackets. Negative values denote concentration. VinDr-CXR endpoints are label-free.

Training design	K	ChestX-ray14	VinDr-CXR
Pooled	–	−4.424 [−4.452, −4.386]	−3.901 [−3.956, −3.864]
Dirichlet	2	−3.641 [−3.843, −3.367]	−2.698 [−2.786, −2.598]
Dirichlet	5	−3.778 [−4.108, −3.444]	−2.776 [−3.538, −1.404]
Dirichlet	20	−3.986 [−4.337, −3.390]	−3.056 [−4.097, −1.334]
View based	2	−2.771 [−3.505, −1.791]	−1.863 [−2.829, −0.883]
View based	5	−3.923 [−4.169, −3.752]	−3.021 [−3.468, −2.422]
View based	20	−3.328 [−4.085, −1.968]	−2.693 [−3.688, −0.912]

On VinDr-CXR, all 21 federated and pooled pairs have lower r_Z and r_C and higher q_1 after alignment. Table 3 shows a negative covariance-rank contrast for every design and seed. The federated $\Delta \log r_C$ is less negative than the pooled contrast in all 18 ChestX-ray14 pairs; differences in exposure and mixed-precision scaling make this comparison descriptive rather than causal.

5 Discussion

Pooled training demonstrates that class-mean alignment can concentrate representations without federation. The same shift on VinDr-CXR shows that the geometry is not confined to the ChestX-ray14 test set. Prediction follows a different pattern: pooled AUROC differences are negative in all seeds, whereas federated changes vary by setting and metric. Effective rank summarizes total variation, but classification may depend on a few directions. The spectra identify concentration, not whether the affected directions are clinical, acquisition-related, or redundant. Since r_Z and r_C weight spectra differently, the estimated concentration also depends on the measure. Together, these findings separate geometric concentration from predictive benefit.

Quadratic attraction to one class mean suppresses within-class variation but does not produce the inter-class or classifier geometry of full neural collapse. Alternative designs target inter-class separation [2,21], angular or relational structure [4,20], or delayed alignment [1]. Evaluating such designs with predictive and spectral endpoints would test whether they retain within-class variation and predictive performance.

The matched design fixes one binary task, one backbone, one alignment weight, and three seeds; VinDr-CXR is label-free; and the client allocations approximate institutional partitions. Broader inference will require prespecified endpoints, labeled external cohorts, and variation in tasks, backbones, and alignment strengths.

6 Conclusion

Class-mean alignment concentrates centered chest X-ray representations in both training designs. The effect persists on VinDr-CXR, while predictive changes depend on context. Spectral and predictive endpoints answer distinct questions.

Data and code availability. We redistribute neither ChestX-ray14 nor VinDr-CXR; dataset access follows NIH and PhysioNet terms [13,12]. Code, aggregate results, figure builders, and checkpoint hashes are at https://github.com/r78zyang/decaf2026-diag.

Ethics and AI assistance. This secondary analysis used de-identified data [13,11] and involved no participant contact. Article 32 of the 2023 Chinese measures identifies qualifying use of lawfully obtained public or anonymized data as eligible for ethics-review exemption [10]. AI-assisted tools supported editing and code review; the authors verified all analyses, references, and text.

References

1. Casado-Diez, M., Dopico-Castro, A., Bolón-Canedo, V., et al.: Closing the alignment-maturity gap in federated prototype learning. arXiv preprint arXiv:2606.02172 (2026). https://doi.org/10.48550/arXiv.2606.02172
2. Dai, Y., Chen, Z., Li, J., et al.: Tackling data heterogeneity in federated learning with class prototypes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 7314–7322 (2023). https://doi.org/10.1609/aaai.v37i6.25891
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
4. Fu, L., Huang, S., Lai, Y., et al.: Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16648–16656 (2025). https://doi.org/10.1609/aaai.v39i16.33829
5. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=YevsQ05DEN7
6. Kairouz, P., McMahan, H.B., Avent, B., et al.: Advances and open problems in federated learning. Foundations and Trends in Machine Learning **14**(1–2), 1–210 (2021). https://doi.org/10.1561/2200000083
7. Li, Z., Shang, X., He, R., Lin, T., Wu, C.: 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. pp. 5296–5306 (2023). https://doi.org/10.1109/ICCV51070.2023.00490
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
9. McMahan, B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR (2017), https://proceedings.mlr.press/v54/mcmahan17a.html
10. National Health Commission of the People's Republic of China: Measures for ethical review of life science and medical research involving humans. National Health Commission website (2023), article 32; accessed 21 July 2026
11. Nguyen, H.Q., Lam, K., Le, L.T., et al.: VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data **9**, 429 (2022). https://doi.org/10.1038/s41597-022-01498-w
12. Nguyen, H.Q., Pham, H.H., Le, T.L., Dao, M., Lam, K.: VinDr-CXR: An open dataset of chest X-rays with radiologist annotations. PhysioNet (2021). https://doi.org/10.13026/3akn-b287, version 1.0.0
13. NIH Clinical Center: NIH Clinical Center provides one of the largest publicly available chest X-ray datasets to the scientific community. NIH Clinical Center website (2017), accessed 21 July 2026
14. Papyan, V., Han, X.Y., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020). https://doi.org/10.1073/pnas.2015509117
15. Qiao, Y., Munir, M.S., Adhikary, A., et al.: MP-FedCL: Multiprototype federated contrastive learning for edge intelligence. IEEE Internet of Things Journal **11**(5), 8604–8623 (2024). https://doi.org/10.1109/JIOT.2023.3320250

16. Rieke, N., Hancox, J., Li, W., et al.: The future of digital health with federated learning. npj Digital Medicine **3**, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1
17. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 15th European Signal Processing Conference. pp. 606–610 (2007). https://doi.org/10.5281/zenodo.40328
18. Tan, Y., Long, G., Liu, L., et al.: FedProto: Federated prototype learning across heterogeneous clients. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8432–8440 (2022). https://doi.org/10.1609/aaai.v36i8.20819
19. Wang, X., Peng, Y., Lu, L., et al.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017). https://doi.org/10.1109/CVPR.2017.369
20. Wu, X., Niu, J., Zhu, G., et al.: From coordinate matching to structural alignment: Rethinking prototype alignment in heterogeneous federated learning. arXiv preprint arXiv:2605.05959 (2026). https://doi.org/10.48550/arXiv.2605.05959
21. Zhou, Y., Qu, X., You, C., et al.: FedSA: A unified representation learning via semantic anchors for prototype-based federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 23009–23017 (2025). https://doi.org/10.1609/aaai.v39i21.34464