Collaborative Synthetic Data Generation for Knowledge Transfer in Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

One-shot federated learning (OSFL) addresses the communication overhead of federated learning by limiting training to a single round, but doing so without sacrificing model quality is non-trivial, particularly when client data distributions diverge. Recent work has addressed this challenge by aggregating client knowledge on the server through the construction of transferable synthetic datasets or distillates. However, most of these methods lack formal privacy guarantees, leaving a gap in jointly achieving low communication, robustness to heterogeneity, and rigorous privacy. We propose FedKT-CSD (Federated Knowledge Transfer via Collaborative Synthetic Data), a framework inspired by neural image compression that closes this gap by leveraging publicly pretrained autoencoders as a shared latent space. Each client encodes its private data in a single forward pass, computes class-conditional latent statistics, and transmits these to the server. The server aggregates these statistics via secure aggregation, adds calibrated differential privacy noise, and decodes a synthetic dataset for training a global model and further downstream tasks. This design provides formal (ε, δ)-differential privacy by construction, while keeping client-side computation and communication lightweight. Despite operating under privacy constraints, FedKT-CSD is competitive with and even outperforms non-private baselines across diverse datasets and heterogeneity settings, and scales to a large number of clients. Our code is available at: `https://github.com/an7123/FedKT-CSD`

1 Introduction

Federated Learning (FL) McMahan et al. (2017) has emerged as a powerful paradigm for training models collaboratively across decentralized data, under the promise of privacy preservation. Nonetheless, the presence of heterogeneous data across clients poses a significant challenge, often leading to slow convergence and suboptimal model performance. A wide range of approaches have been proposed to address this, from personalized federated learning (pFL) Luo & Wu (2022); Zhang et al. (2021); Dinh et al. (2020); Sun et al. (2021); Li et al. (2021); Collins et al. (2021) to data and representation sharing strategies Zhao et al. (2018); Zhu et al. (2021); Zhang et al. (2022b); Luo et al. (2024); Yang et al. (2023); Hoefler et al. (2025). Despite their effectiveness, these methods typically require many communication rounds which can be problematic for real-world cross-device deployment.

A particularly attractive direction for communication constrained applications is one-shot federated learning (OSFL), where clients communicate with the server exactly once. Recent OSFL methods leverage the aforementioned data sharing principle, showing the efficacy in improving one-shot performance. Methods such as FedD3 Song et al. (2023), FedSD2C Zhang et al. (2024), FedCVAE Heinbaugh et al. (2023), DENSE Zhang et al. (2022a), CoBoosting Dai et al. (2024) and Beitollahi et al. (2025) approach OSFL by having each client train a local generative model or distill its data, then transmit the result to the server, which aggregates the contributions into a shared synthetic dataset or ensemble. These approaches thus eliminate repeated parameter exchanges, naturally support heterogeneous architectures, and can produce datasets that can be reused for pretraining, distillation, or personalization.

However, existing OSFL methods face a fundamental limitation: few provide formal privacy guarantees as a built-in mechanism with regards to preserving utility. Privacy leakage occurs because clients transmit trained model parameters, generator weights, or distilled data, where information about individual training records can be leaked. Retrofitting differential privacy (DP) Dwork & Roth (2014) onto these approaches is difficult; the high dimensionality of model parameters inflates sensitivity, DP noise degrades already-fragile local models, and multi-step training complicates privacy accounting. Meanwhile, training local generators and models such as in Heinbaugh et al. (2023); Zhang et al. (2022a); Dai et al. (2024) can also heavily increase computational cost on clients.

This leads to a direct question: *Can we devise a framework which leverages the benefits of representation sharing in an OSFL setting with formal privacy guarantees and lightweight client computation?*

We observe that modern pretrained autoencoders, originally developed for neural image compression Chen et al. (2024); Ballé et al. (2018), provide exactly the structure needed to answer this question. These models map images to compact latent vectors and back with high fidelity, using publicly available weights that require no adaptation. The key insight is that, given such a shared public encoder, each client can encode its private images in a single forward pass and compute simple per-class statistics in the latent space. These statistics can be aggregated across clients using secure aggregation, which we implement such that partitioning data across any number of clients produces exactly the same as if data were aggregated centrally. Moreover, the latent vectors are also low-dimensional and have bounded sensitivity, making them ideal targets for calibrated Gaussian noise under differential privacy Dwork (2006).

Building on this insight, we propose **FedKT-CSD** (Federated Knowledge Transfer via Collaborative Synthetic Data). Each client receives a frozen, publicly pretrained autoencoder and encodes its private images via a single forward pass with no optimization or training on-device. The client computes class-conditional latent statistics, protects them under differential privacy, and transmits only these lightweight summaries to the server. The server aggregates the noisy statistics across all clients, recovers per-class Gaussian distributions in latent space, samples from them, and decodes the samples into a synthetic image dataset. This dataset can then be used to train a global classifier, a feature extractor for personalized FL, or auxiliary data for multi-round methods.

Our approach achieves three critical objectives simultaneously: (1) privacy by design, since the autoencoder is public and frozen and only DP-protected class-conditional statistics are transmitted; (2) communication efficiency, as the approach requires a single communication round where each client uploads only statistics rather than model parameters; and (3) computational efficiency, since clients perform only a forward pass through the encoder with no on-device training, optimization, or gradient computation.

Our contributions are:

- **One-shot DP synthetic data generation.** We show that deep compression autoencoders provide a shared latent space in which class-conditional statistics can be computed with differential privacy. This enables a one-shot FL pipeline that generates synthetic data with formal (ε, δ)-DP guarantees, a single communication round, and no on-device training.

- **Competitive with SOTA under DP.** Despite operating under DP, FedKT-CSD is competitive and outperforms existing one-shot FL baselines across diverse datasets and heterogeneity settings. Moreover, our method is also invariant to both the degree of heterogeneity and the number of clients, properties that no existing OSFL method possesses.

- **Versatile downstream use.** The DP synthetic dataset is not limited to server-side training. We demonstrate its effectiveness for downstream personalized FL, where it serves as pretraining data and achieves improvements over multi-round personalized FL baselines.

- **Practical efficiency.** Each client uploads payloads in the regime of Kilobytes and requires seconds of computation with no GPU training. This makes FedKT-CSD deployable in bandwidth- and compute-constrained settings where existing methods are impractical.

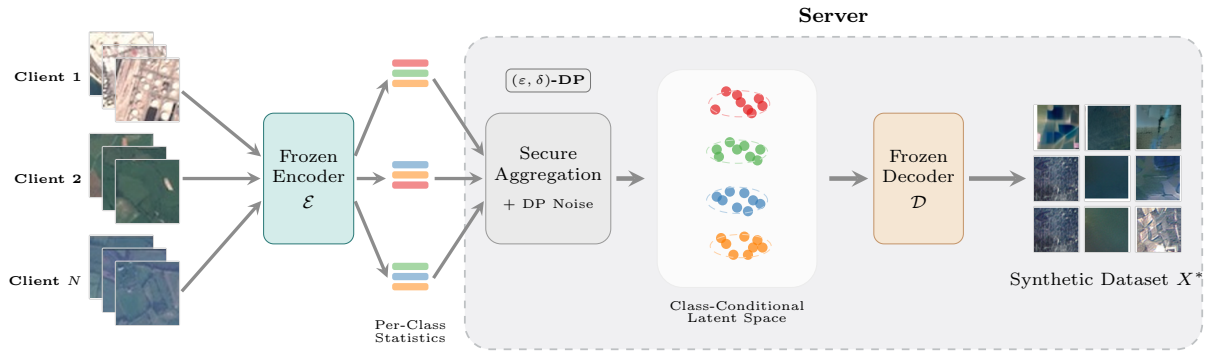


Figure 1: Overview of FedKT-CSD. Each client holds private images from a heterogeneous subset of classes. A shared frozen encoder maps images to compact latent vectors, from which per-class statistics (colored bars) are computed locally. These lightweight summaries are transmitted via secure aggregation and protected with calibrated DP noise on the server. The server recovers class-conditional Gaussian distributions in the latent space (colored clusters), samples new latent vectors, and decodes them into a synthetic dataset X^* via the frozen decoder.

2 Related Works and Preliminaries

One-Shot Federated Learning. One-shot FL methods can be organized by how clients communicate with the server. One family transmits local models or generators, from which the server extracts knowledge via distillation. DENSE Zhang et al. (2022a) and Co-Boosting Dai et al. (2024) follow this approach, aggregating client model ensembles through data-free knowledge distillation. Both evaluate privacy empirically through membership inference but provide no formal guarantees. A second family transmits data, or data-derived features, ranging from distilled datasets, synthetic samples, or parametric distributions. FedD3 Song et al. (2023) constructs compact client summaries via dataset distillation. FedSD2C Zhang et al. (2024) optimizes synthetic distillates through a pretrained autoencoder to minimize information loss relative to the original data. Similarly, Heinbaugh et al. (2023) shares locally trained generators to avoid model aggregation, though without privacy analysis. Most closely related to our work is FedPFT Beitollahi et al. (2025), which also encodes images through a frozen pretrained model and transmits class-conditional Gaussian mixtures in a single round. Our approach differs in that we aggregate additive statistics rather than per-client mixture models, enabling secure aggregation, exact partition invariance, and cleaner DP accounting. In the FedPFT privacy model, noise is added per-client to high-dimensional GMM parameters, which does not benefit from secure aggregation and is likely to degrade at meaningful privacy budgets. Moreover, we use an autoencoder rather than a discriminative feature extractor, producing a reusable pixel-space synthetic dataset rather than features tied to a specific classifier head, which is more useful for downstream tasks. We discuss the precise differences further in the appendix.

Data and Representation Sharing. Underpinning the recent works in OSFL are works on data sharing. To improve transferability, recent work explores sharing auxiliary information beyond parameters. FedGen Zhu et al. (2021) and FedFTG Zhang et al. (2022b) use knowledge distillation from client logits to train server-side generators or refine global models. CCVR Luo et al. (2024) and FedFed Yang et al. (2023) share representation statistics to calibrate classifiers or align features. In addition, Hoefler et al. (2025) use data derived from XAI-attribution methods, and Hoefler et al. (2024) uses diffusion models to account for classifier drift. More recently, foundation models enable high-quality synthetic data generation: Pre-Text Hou et al. (2024), dp-prompt Utpala et al. (2023), and dp-kde Nagesh et al. (2025) leverage LLMs in the text domain. These approaches typically require large pretrained models and may incur significant computational or privacy costs.

Algorithm 1 FedKT-CSD Algorithm

Input: Frozen encoder $\mathcal{E}$, decoder $\mathcal{D}$; N clients with labeled data; privacy budget (ε, δ); clipping radius R; synthetic samples per class N_c

Output: DP synthetic dataset X^*

Client side (each client i, in parallel, no training)

1. **for** each image $(x_{i,j}, y_{i,j}) \in \mathcal{D}_i$ **do**
   a. Encode: $z_{i,j} \leftarrow \mathcal{E}(x_{i,j})$
   b. Clip to bounded norm: $\bar{z}_{i,j} \leftarrow \mathrm{Clip}_R(z_{i,j})$
2. **for** each class c present in client i's data **do**
   a. Sum of latents: $\mathbf{m}_c^i \leftarrow \sum_j \bar{z}_{i,j}$
   b. Sum of outer products: $\mathbf{M}_c^i \leftarrow \sum_j \bar{z}_{i,j}\bar{z}_{i,j}^\top$
   c. Sample count: $n_c^i \leftarrow$ number of class-c images
3. Upload $\{(\mathbf{m}_c^i, \mathbf{M}_c^i, n_c^i)\}_c$ via secure aggregation

Server side (aggregate, add noise, generate)

5. Aggregate across clients: $\mathbf{S}_{\mu,c} \leftarrow \sum_i \mathbf{m}_c^i, \quad \mathbf{S}_{\Sigma,c} \leftarrow \sum_i \mathbf{M}_c^i, \quad n_c \leftarrow \sum_i n_c^i$
6. Add calibrated Gaussian noise to each sum (Prop. 4.1)
7. **for** each class $c = 1, \ldots, K$ **do**
   a. Recover noisy mean: $\boldsymbol{\mu}_c \leftarrow$ noisy latent sum $/\, n_c$
   b. Recover covariance: second moment $/\, n_c$ minus outer product of mean
   c. Apply bias correction and PSD projection
   d. Sample N_c latents from $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$
   e. Decode to images: $x_c^{(k)} \leftarrow \mathcal{D}(z_c^{(k)})$
8. **Return** $X^* = \{(x_c^{(k)}, c)\}$ for downstream training

3 Methodology

3.1 Problem Setting

We consider one-shot federated learning (OSFL). N clients hold private datasets $\mathcal{D}_i = \{(x_{i,j}, y_{i,j})\}_{j=1}^{n_i}$ with $y_{i,j} \in \{1, \ldots, K\}$, drawn from heterogeneous distributions. The goal is to produce a global model f_θ minimizing

$$\min_\theta \sum_{i=1}^{N} \frac{n_i}{\sum_j n_j} \mathcal{L}_{\mathcal{D}_i}(\theta), \tag{1}$$

subject to (i) a **single communication round** and (ii) **record-level differential privacy**. Our key observation is that statistics of latent representations are additive: class-conditional sums computed locally combine across clients without information loss, so each client transmits only lightweight summary statistics from which the server generates a DP synthetic dataset X^* (Figure 1). We follow the standard conditional-generation setting where class labels are treated as public information. The privacy guarantee therefore protects the private feature content of each record, not its class membership although we outline how to achieve this in the appendix.

3.2 Pretrained Autoencoder and Encoding

We use a publicly pretrained autoencoder with encoder $\mathcal{E}\colon \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$ and decoder $\mathcal{D}\colon \mathbb{R}^d \to \mathbb{R}^{H\times W\times C}$, both frozen throughout.

Each client i encodes its images via $z_{i,j} = \mathcal{E}(x_{i,j})$, a single forward pass with no optimization or training on-device. After clipping each latent to ℓ_2-norm at most R (Section 4.1), the client groups its samples by class label. Let $\mathcal{I}_{i,c} = \{j : y_{i,j} = c\}$ denote the index set of client i's samples belonging to class c. For each class it holds, the client computes three quantities:

$$\mathbf{m}_c^i = \sum_{j\in\mathcal{I}_{i,c}} \bar{z}_{i,j}, \qquad \mathbf{M}_c^i = \sum_{j\in\mathcal{I}_{i,c}} \bar{z}_{i,j}\,\bar{z}_{i,j}^\top, \qquad n_c^i = |\mathcal{I}_{i,c}|, \tag{2}$$

where $\bar{z}_{i,j} = \mathrm{Clip}_R(z_{i,j})$ is the clipped latent vector. Here $\mathbf{m}_c^i \in \mathbb{R}^d$ is the sum of clipped latents for class c (not the mean), $\mathbf{M}_c^i \in \mathbb{R}^{d\times d}$ is the sum of their outer products, and n_c^i is the number of class-c samples held by client i. The triple $(\mathbf{m}_c^i, \mathbf{M}_c^i, n_c^i)$ is all that leaves the client. We transmit sums rather than means because sums are additive: they can be aggregated across clients without knowing per-client sample counts, compose naturally with secure aggregation (Section 3.3), and have bounded per-record sensitivity for DP.

3.3 Aggregation and Noise Injection

The server collects the per-client triples and aggregates them via secure aggregation, a cryptographic protocol that computes element-wise sums across clients while revealing only the totals to the server, never individual contributions (we describe secure aggregation further in the appendix). The resulting global quantities are:

$$\mathbf{S}_{\mu,c} = \sum_{i=1}^{N} \mathbf{m}_c^i \in \mathbb{R}^d, \qquad \mathbf{S}_{\Sigma,c} = \sum_{i=1}^{N} \mathbf{M}_c^i \in \mathbb{R}^{d\times d}, \qquad n_c = \sum_{i=1}^{N} n_c^i. \tag{3}$$

Because these are simple sums, the result is identical to what would be computed if all class-c data resided on a single machine; the partitioning across clients has no effect. Gaussian noise calibrated to the sensitivity bounds (Proposition 4.1) is then added to the aggregated sums:

$$\widetilde{\mathbf{S}}_{\mu,c} = \mathbf{S}_{\mu,c} + \eta_\mu, \qquad \widetilde{\mathbf{S}}_{\Sigma,c} = \mathbf{S}_{\Sigma,c} + \eta_\Sigma, \tag{4}$$

where $\eta_\mu \sim \mathcal{N}(0, \sigma_\mu^2 I_d)$ and $\eta_\Sigma \sim \mathcal{N}(0, \sigma_\Sigma^2 I_{d\times d})$. After this step, all subsequent operations (normalization, sampling, decoding, model training) are deterministic functions of the noisy sums and incur no additional privacy cost.

3.4 Post-Processing

Given the noisy sums $(\widetilde{\mathbf{S}}_{\mu,c}, \widetilde{\mathbf{S}}_{\Sigma,c}, n_c)$, the server recovers class-conditional statistics. The DP-protected mean $\boldsymbol{\mu}_c^{\mathrm{DP}} \in \mathbb{R}^d$ and raw covariance estimate $\widehat{\boldsymbol{\Sigma}}_c \in \mathbb{R}^{d\times d}$ are:

$$\boldsymbol{\mu}_c^{\mathrm{DP}} = \frac{\widetilde{\mathbf{S}}_{\mu,c}}{n_c}, \qquad \widehat{\boldsymbol{\Sigma}}_c = \frac{\widetilde{\mathbf{S}}_{\Sigma,c}}{n_c} - \boldsymbol{\mu}_c^{\mathrm{DP}}\,(\boldsymbol{\mu}_c^{\mathrm{DP}})^\top. \tag{5}$$

The first term $\widetilde{\mathbf{S}}_{\Sigma,c}/n_c$ is the noisy second moment; subtracting the outer product of the mean yields the covariance. Two corrections are needed before $\widehat{\boldsymbol{\Sigma}}_c$ can be used for sampling:

Bias correction. Because $\boldsymbol{\mu}_c^{\mathrm{DP}}$ is itself noisy, the subtracted outer product is biased upward: $\mathbb{E}[\boldsymbol{\mu}_c^{\mathrm{DP}}(\boldsymbol{\mu}_c^{\mathrm{DP}})^\top] = \boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top + (\sigma_\mu/n_c)^2 I_d$, where the second term is the variance of the noise in the mean. This increases the subtracted quantity, *decreasing* the covariance estimate along the diagonal. We correct analytically:

$$\widehat{\boldsymbol{\Sigma}}_c \;\leftarrow\; \widehat{\boldsymbol{\Sigma}}_c + \frac{\sigma_\mu^2}{n_c^2}\, I_d. \tag{6}$$

PSD projection. The additive noise on $\widetilde{\mathbf{S}}_{\Sigma,c}$ can result in $\widehat{\mathbf{\Sigma}}_c$ being indefinite (some eigenvalues negative), which would prevent sampling. We project onto the positive semidefinite cone and compute the eigendecomposition $\widehat{\mathbf{\Sigma}}_c = Q\Lambda Q^\top$ (with Q orthogonal and Λ diagonal) and clamp all eigenvalues to a minimum threshold $\tau > 0$:

$$\mathbf{\Sigma}_c^{\mathrm{DP}} = Q\,\max(\Lambda,\, \tau I)\,Q^\top, \tag{7}$$

where τ is small (e.g., 10^{-6}). The result $\mathbf{\Sigma}_c^{\mathrm{DP}}$ is the final DP-protected covariance used for generation. Both corrections are deterministic post-processing of the noisy sums and incur no additional privacy cost.

3.5 Synthetic Data Generation and Downstream Use

The server now possesses, for each class c, a DP-protected mean $\boldsymbol{\mu}_c^{\mathrm{DP}} \in \mathbb{R}^d$ and covariance $\mathbf{\Sigma}_c^{\mathrm{DP}} \in \mathbb{R}^{d\times d}$ that together parameterize a class-conditional Gaussian in the autoencoder's latent space. For each class c, N_c latent vectors are sampled from this Gaussian and decoded back to pixel space via the frozen decoder $\mathcal{D}$:

$$z_c^{(k)} \sim \mathcal{N}\big(\boldsymbol{\mu}_c^{\mathrm{DP}},\, \mathbf{\Sigma}_c^{\mathrm{DP}}\big), \qquad x_c^{(k)} = \mathcal{D}(z_c^{(k)}), \qquad k = 1, \ldots, N_c. \tag{8}$$

The synthetic dataset $X^* = \bigcup_{c=1}^{K} \{(x_c^{(k)}, c)\}_{k=1}^{N_c}$ aggregates decoded images across all K classes. Since X^* is derived entirely from the DP-protected statistics, it can be generated in any quantity, producing more samples is just additional sampling and decoding, at no additional privacy cost.

In our experiments we train a neural network on X^* via standard ERM:

$$\theta^* = \arg\min_{\theta} \frac{1}{|X^*|} \sum_{(x,y)\in X^*} \ell\big(f_\theta(x),\, y\big). \tag{9}$$

X^* is equally suitable as pretraining data for multi-round FL, a distillation corpus, or input to any other learning pipeline. We refer do Figure 1 for a visualization of our method.

4 Privacy Analysis

The DP analysis is straightforward compared to iterative FL methods. Statistics are released exactly once, sensitivity is determined by a public clipping threshold, and all downstream operations are post-processing. This DP framework does not protect class membership and we assume the label space is public knowledge. We show how to privatize class membership in the supplementary materials.

4.1 Clipping

Before computing statistics, each client clips its encoded latents to ℓ_2-norm to R:

$$\bar{z}_{i,j} = z_{i,j} \cdot \min\Big(1,\, \frac{R}{\|z_{i,j}\|_2}\Big). \tag{10}$$

Clipping ensures that each individual record's contribution to the sums $\mathbf{S}_{\mu,c}$ and $\mathbf{S}_{\Sigma,c}$ is bounded, which is necessary for calibrating the DP noise. The threshold R is set data-independently from the autoencoder prior: under $p(z) = \mathcal{N}(0, I_d)$, the norm $\|z\|_2$ concentrates near $\sqrt{d}$, so $R = \alpha\sqrt{d}$ with a moderate α (e.g., $\alpha = 3$) covers the vast majority of encoded latents without inspecting any private data.

4.2 Sensitivity

Under the public-label assumption above, we adopt *within-class replacement* adjacency. Adjacent datasets $\mathcal{D} \sim \mathcal{D}'$ differ by replacing one class-c record with another class-c record, leaving all counts n_c unchanged. Under this definition, replacing a clipped latent $\bar{z}$ (with $\|\bar{z}\|_2 \le R$) by $\bar{z}'$ (with $\|\bar{z}'\|_2 \le R$) changes the sum $\mathbf{S}_{\mu,c}$ by $\bar{z}' - \bar{z}$ and the second-moment sum $\mathbf{S}_{\Sigma,c}$ by $\bar{z}'(\bar{z}')^\top - \bar{z}\bar{z}^\top$. The ℓ_2-sensitivities are therefore:

$$\Delta_\mu = \sup_{\mathcal{D}\sim\mathcal{D}'} \|\mathbf{S}_{\mu,c}(\mathcal{D}) - \mathbf{S}_{\mu,c}(\mathcal{D}')\|_2 \le 2R, \qquad \Delta_\Sigma = \sup_{\mathcal{D}\sim\mathcal{D}'} \|\mathbf{S}_{\Sigma,c}(\mathcal{D}) - \mathbf{S}_{\Sigma,c}(\mathcal{D}')\|_F \le 2R^2, \tag{11}$$

by the triangle inequality and the norm bound on clipped latents.

4.3 Privacy Guarantee

The mechanism results in two quantities per class: the noisy mean sum $\widetilde{\mathbf{S}}_{\mu,c}$ and the noisy second-moment sum $\widetilde{\mathbf{S}}_{\Sigma,c}$, each protected by the Gaussian mechanism. We split the total privacy budget (ε, δ) equally between them. We use the *analytic Gaussian mechanism* Balle & Wang (2018) rather than the textbook closed-form calibration because it provides a valid and tighter calibration for all $\varepsilon > 0$, including our experimental regime.

For a query $q(\mathcal{D})$ with ℓ_2-sensitivity Δ, we release

$$\widetilde{q}(\mathcal{D}) = q(\mathcal{D}) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I), \tag{12}$$

where σ is calibrated by the analytic Gaussian mechanism. In our case, this gives

$$\widetilde{\mathbf{S}}_{\mu,c} = \mathbf{S}_{\mu,c} + \eta_{\mu,c}, \qquad \widetilde{\mathbf{S}}_{\Sigma,c} = \mathbf{S}_{\Sigma,c} + \eta_{\Sigma,c}, \tag{13}$$

where $\eta_{\mu,c} \sim \mathcal{N}(0, \sigma_\mu^2 I_d)$ and, viewing $\mathbf{S}_{\Sigma,c} \in \mathbb{R}^{d\times d}$ as a vector in $\mathbb{R}^{d^2}$, $\eta_{\Sigma,c}$ is isotropic Gaussian noise in that ambient space with variance σ_Σ^2 per coordinate.

Proposition 4.1 (FedKT-CSD Differential Privacy)**.** *Under the notation above, let $\varepsilon' = \varepsilon/2$ and $\delta' = \delta/2$. Set the noise scales σ_μ and σ_Σ to be the smallest positive values such that*

$$\Phi\left(\frac{\Delta}{2\sigma} - \frac{\varepsilon'\sigma}{\Delta}\right) - e^{\varepsilon'}\Phi\left(-\frac{\Delta}{2\sigma} - \frac{\varepsilon'\sigma}{\Delta}\right) \le \delta', \tag{14}$$

where Φ is the standard normal CDF, with $\Delta = \Delta_\mu$ for σ_μ and $\Delta = \Delta_\Sigma$ for σ_Σ. Then, for a fixed class c, the joint disclosure of $(\widetilde{\mathbf{S}}_{\mu,c}, \widetilde{\mathbf{S}}_{\Sigma,c})$ satisfies (ε, δ)-differential privacy at the record level.

Proof. By the analytic Gaussian mechanism Balle & Wang (2018), a query with ℓ_2-sensitivity Δ satisfies (ε', δ')-DP when Gaussian noise with standard deviation σ is calibrated according to Eq. equation 14. Applying this separately to the two queries with sensitivities $\Delta_\mu = 2R$ and $\Delta_\Sigma = 2R^2$ yields $(\varepsilon/2, \delta/2)$-DP for each release. By basic composition, the joint release satisfies $(\varepsilon/2{+}\varepsilon/2,\ \delta/2{+}\delta/2) = (\varepsilon, \delta)$-DP. □

All subsequent operations such as normalization (Eq. 5), bias correction (Eq. 6), PSD projection (Eq. 7), sampling, decoding, and model training are deterministic functions of the released noisy sums and preserve (ε, δ)-DP by the post-processing theorem (Dwork & Roth, 2014). The effective noise in the mean scales as σ_μ/n_c, so larger per-class counts directly improve utility. Within-class replacement protects the *values* of individual records (i.e., which specific image a client contributes to class c) but does not protect class membership. The alternative add/remove adjacency, which does hide class membership, would reduce sensitivity to R (resp. R^2) since only one term is added or removed rather than swapped, but requires additional care in handling the variable count n_c. However, class membership can be protected via secure aggregation and shuffling methods; see the appendix and Kairouz et al. (2021). In addition, we discuss making the class count DP in the appendix.

5 Experimental Results

Datasets. We evaluate on four image classification benchmarks chosen to cover diverse visual domains: ImageNette (10 classes, 9.5k ImageNet subset) and CIFAR-100 (100 classes, 50k natural images) as standard benchmarks for image classification, and EuroSAT (10 classes, 27k satellite images), and BloodMNIST (8 classes, 12k microscopy images) since these constitute real-world and out-of-distribution datasets for the pretrained VAE encoder-decoder pair. This selection tests whether the pretrained autoencoder, originally trained on natural images, generalizes to other domains.

Heterogeneity simulation. We follow Zhang et al. (2025) for experimental setups. The *practical* setting draws per-client label distributions from a Dirichlet distribution with concentration parameter α where smaller α means more heterogeneity. The *pathological* setting assigns each client a disjoint subset of classes (2 classes per client for 10- and 8-class datasets; 10 for CIFAR-100). Together these cover the spectrum from moderate label imbalance to complete class separation.

Table 1: Comparison of OSFL methods. Path. = pathological split (2 classes/client for 10- and 8-class datasets; 10 classes/client for CIFAR-100). Best in **bold**.

	EuroSAT (10 cls, 27k)			**CIFAR-100** (100 cls, 50k)			**ImageNette** (10 cls, 9.5k)			**BloodMNIST** (8 cls, 12k)		
Method	α=0.1	α=0.05	Path.	α=0.1	α=0.05	Path.	α=0.1	α=0.05	Path.	α=0.1	α=0.05	Path.
FedSD2C	75.83	68.47	54.62	27.19	23.62	20.99	57.24	48.36	35.82	69.52	61.83	65.47
DENSE	62.38	53.71	40.25	10.61	5.55	3.82	42.17	31.54	21.63	58.46	48.72	54.83
CoBoosting	58.62	49.83	36.47	16.92	15.70	14.95	39.36	28.53	19.27	54.82	45.63	51.47
FedD3	42.35	38.72	31.46	1.82	1.74	1.61	24.18	20.65	16.83	48.36	43.85	50.27
FedCVAE	40.83	37.45	30.62	1.64	1.65	1.55	25.47	21.38	17.52	50.15	45.28	51.83
Ours (DP)	**80.96**	**80.74**	**80.46**	**29.71**	**28.69**	**28.62**	**62.23**	**61.98**	**62.46**	**75.03**	**74.84**	**75.19**

Implementation. We follow Zhang et al. (2025) in their implementation. Default client number is 20 clients with full participation. The autoencoder is DC-AE f32c32 (Chen et al., 2024). Privacy parameters are ε=10, δ=10^{-5} unless stated otherwise. Downstream classifier is ResNet-18 trained on synthetic data only and evaluated on real test data of the corresponding dataset. All OSFL baselines use a ResNet-18 unless otherwise specified.

5.1 One-Shot FL Performance Comparison

Tables 1 compare FedKT-CSD against five one-shot FL baselines (FedSD2C, DENSE, CoBoosting, FedD3, and FedCVAE) across four datasets and three heterogeneity settings. All baselines operate without differential privacy, while FedKT-CSD satisfies (ε=10, δ=10^{-5})-DP. Despite this stricter threat model, FedKT-CSD outperforms every baseline on every dataset and heterogeneity setting.

The reason lies in how each method handles the heterogeneity-communication tradeoff. The baselines train local models (generators, classifiers, or distilled representations) on each client's skewed data, then aggregate these models on the server. When data is heterogeneous, local models are biased toward the client's label subset, and single-round aggregation cannot fully correct this bias. This effect is visible in the tables. Every baseline degrades by 20–40% between α=0.1 and the pathological setting, because pathological splits produce heavily biased local models.

FedKT-CSD sidesteps this entirely. Because we aggregate per-class statistics rather than model parameters, the server reconstructs the same global class-conditional distributions regardless of how labels are distributed across clients. The three heterogeneity columns within each dataset are identical up to DP noise, a property that no model-aggregation approach can match. This heterogeneity invariance is a structural guarantee of the additive aggregation scheme.

Furthermore, the performance gap is widest on datasets where baseline local models are most fragile. On CIFAR-100, where pathological splits give each client only 10 of 100 classes, baselines collapse below 21% while FedKT-CSD maintains ∼29%. On EuroSAT and BloodMNIST, which have fewer classes and more samples per class, baselines fare better but FedKT-CSD still leads by 7–15 points. The pattern is consistent: the harder the heterogeneity makes local model training, the larger our advantage.

5.2 Communication and Compute

Table 2 profiles the client-side cost of FedKT-CSD for four autoencoder configurations. All timings are wall-clock measurements on a single NVIDIA RTX 5090 processing 1000 images. The pipeline has three stages: encoding, DP statistics computation and sampling, and decoding.

With our default configuration (DC-AE f32c32 at 64×64), the entire pipeline completes in 1.33 seconds, dominated by the encoder and decoder forward passes (0.47s and 0.85s respectively). The statistics and sampling step is negligible at 3ms, since it involves only vector sums and matrix operations in $\mathbb{R}^{128}$. The lightweight TAESD codec at 32×32 finishes in under 0.1 seconds total, making it viable even on CPU-only clients, though at the cost of lower downstream accuracy (Table 5).

Table 2: **Client-side cost** (1000 images, 10 classes, RTX 5090). Zero training: encode → DP stats → sample → decode.

	TAESD 32×32	**DC-AE f32c32** 64×64	**DC-AE f64c128** 64×64	**DC-AE f32c32** 128×128
Enc./Dec. params	1.2 M / 1.2 M	323 M / 323 M	677 M / 677 M	323 M / 323 M
Latent dim d	64	128	128	512
Model weights	9 MB	1.2 GB	2.6 GB	1.2 GB
Encode (s)	0.020	0.474	0.473	1.616
DP stats + sample (s)	0.009	0.003	0.002	0.006
Decode (s)	0.059	0.850	0.939	3.433
Total (s)	0.09	1.33	1.41	5.06
Peak GPU (encode)	333 MB	3.5 GB	4.8 GB	10.1 GB
Client RAM	21 MB	1.3 GB	2.6 GB	1.4 GB
Upload (1×)	83.8 KB	327.5 KB	327.5 KB	5.0 MB

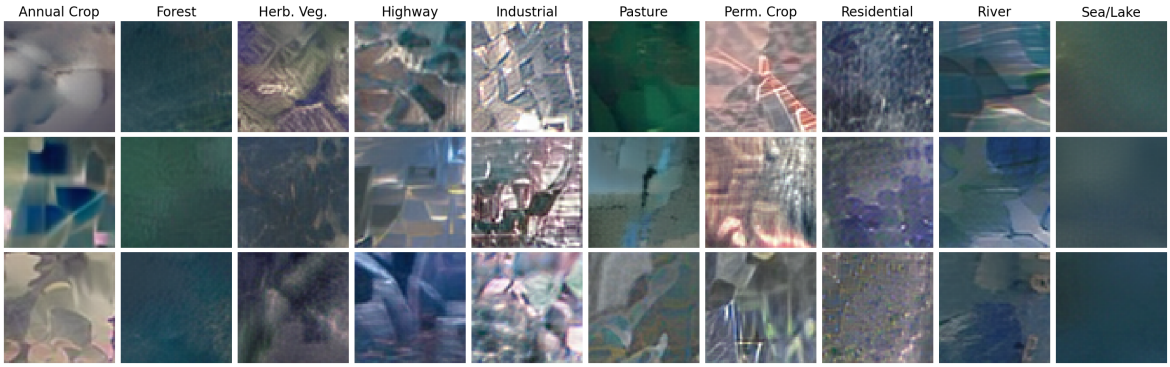


Figure 2: Synthetic EuroSAT images generated by FedKT-CSD (ε=10, δ=10^{-5}). Three random samples per class, decoded from DP-protected class-conditional Gaussians via the frozen DC-AE decoder.

Upload cost is the total payload a client transmits to the server. A client sends one mean vector (d floats) and one symmetric covariance matrix ($d(d+1)/2$ floats) per class. With d=128 and 10 classes this amounts to 327.5 KB per client — approximately two orders of magnitude smaller than a single FedAvg model upload (a ResNet-18 is ∼45 MB). The high-resolution DC-AE at 128×128 incurs a substantially larger upload of 5.0 MB due to the d=512 latent dimension, which also degrades DP utility (Table 5), making it unattractive on both axes. Peak GPU memory during encoding is 3.5 GB for DC-AE f32c32, well within the capacity of consumer hardware. For deployment on edge devices without a GPU, the TAESD configuration requires only 333 MB of GPU memory and 21 MB of client RAM.

5.3 Visualization

Figure 2 shows samples of synthetic EuroSAT images generated from the DP-protected class-conditional Gaussians. The samples are not entirely photorealistic, because each class is modeled as a single Gaussian in a 128-dimensional latent space, yet the decoder produces images that capture the dominant visual modes of each class rather than fine-grained spatial detail. Moreover, class-discriminative structure is clearly visible. Forests and pastures produce uniform green and blue tones, residential areas contain grid-like patterns with visible rooftops, highways exhibit linear features, and industrial regions display angular structures. Annual and permanent crops are distinguishable by texture regularity, while sea/lake samples show consistent dark blue hues. These visual differences align with what a downstream classifier needs to separate the classes, which explains why the synthetic data achieves ∼80% accuracy on EuroSAT despite the coarse appearance of individual images. The key insight is that classification does not require photorealism; it requires that the

Table 3: **Comparison with DP synthetic data generation methods.** All methods: (ε=10, δ=10^{-5})-DP. ResNet-18 trained on synthetic data, tested on real data.

Method	Setting	EuroSAT	CIFAR-100	ImageNette	BloodMNIST
Centralized DP synthesis					
DP-MERF	Central	33.70	3.60	15.81	49.38
DP-NTK	Central	38.60	3.60	12.42	50.12
DP-Kernel	Central	47.90	5.70	18.25	61.39
DP-LoRA LDM	Central	**84.60**	**35.20**	58.37	**80.06**
Federated DP synthesis (20 clients)					
DP-Kernel	Fed.	28.35	7.16	16.42	52.62
DP-MERF	Fed.	22.41	4.85	11.93	39.62
DP-NTK	Fed.	19.83	4.52	11.27	48.75
DP-LoRA LDM	Fed.	31.56	9.24	22.18	19.76
Ours					
TAESD	Fed.	76.27	27.08	48.63	74.59
DC-AE f32c32	Fed.	80.68	29.34	**61.10**	74.78

synthetic distribution preserves the inter-class separation present in the real data, which the Gaussian model in latent space achieves effectively. We show more images of other dataset in the appendix.

5.4 Comparison with other DP Synthetic Data Methods

Since we generate a DP synthetic dataset, we also compare FedKT-CSD against established DP synthetic image generation methods (see Gong et al. (2025) for a comprehensive overview of DP methods for images) at ε=10, in both centralized and federated settings in Table 3. For the centralized setup we run the methods using the implementations in Gong et al. (2025). For the federated setup we simply run each method on the local client split and share the generated data to train a model on the server.

The kernel-based methods (DP-Kernel Jiang et al. (2023), DP-MERF Harder et al. (2021), DP-NTK Yang et al. (2024)) operate by matching noisy statistics of random feature embeddings between real and synthetic data. The high-dimensional matching objectives require substantial noise at practical ε values. FedKT-CSD outperforms all methods in the federated settings. This is because our pipeline sidesteps iterative optimization entirely. The pretrained DC-AE encoder provides a rich, low-dimensional embedding (d=128) in which simple Gaussian modeling already captures class structure well.

Furthermore, DP-LoRA LDM Tsai et al. (2025) takes a different approach, fine-tuning a latent diffusion model pretrained on the full ImageNet dataset using DP-SGD. In the centralized setting this is the strongest competitor, since the diffusion model can generate high-fidelity images and the full dataset provides enough gradient signal to absorb DP-SGD noise. However, the federated variant collapses on several datasets (e.g., 20% on BloodMNIST). This is expected, as DP-SGD noise scales with the number of gradient steps and the sensitivity of each step, so when local datasets shrink in a federated split, the signal-to-noise ratio becomes prohibitive. FedKT-CSD does not suffer from this because it never performs gradient-based optimization on private data. The noise is injected once into low-dimensional additive statistics, and the total noise magnitude is independent of how data is partitioned across clients.

5.5 Downstream Tasks

A key advantage of our approach is that the synthetic dataset X^* consists of actual images, not features tied to a specific model. Methods such as FedPFT Beitollahi et al. (2025) generate synthetic representations in a fixed feature space, meaning any downstream model must use the same feature extractor. By contrast, X^* is model-agnostic and can train any architecture, serve as a pretraining corpus, or be combined with local data for fine-tuning.

Table 4: Downstream personalized FL comparison with multi-round pFL methods. All methods use ResNet-18 with 20 clients and full participation. Baselines run 50 communication rounds; FedKT-CSD uses one round (synthetic pretraining) followed by local classifier fine-tuning.

	Pathological Heterogeneity		Practical Heterogeneity	
	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
FedAvg McMahan et al. (2017)	86.05	44.88	79.68	32.77
Per-FedAvg Fallah et al. (2020)	90.12	55.98	87.93	45.38
pFedMe Dinh et al. (2020)	91.21	59.88	89.78	48.12
FedAMP Huang et al. (2021)	89.85	63.32	89.14	49.68
FedPHP Li et al. (2021)	89.12	61.45	87.69	51.21
FedFomo Zhang et al. (2021)	90.78	61.49	87.54	46.21
APPLE Luo & Wu (2022)	90.02	64.84	88.57	52.14
PartialFed Sun et al. (2021)	88.95	60.41	87.29	49.85
FedKT-CSD	**92.82**	**73.44**	**88.78**	**59.89**

We demonstrate this flexibility by using X^* as a pretraining resource for personalized federated learning, following the pFL-Bench protocol Zhang et al. (2025). We use 20 clients with full participation and a ResNet-18 backbone. Baselines train via multi-round optimization (50 rounds). For FedKT-CSD, the server pretrains ResNet-18 on X^*, distributes it, and each client fine-tunes only the classification layer locally. Crucially, this ResNet-18 is entirely separate from the autoencoder that generated X^*, i.e., the synthetic images transfer across model families.

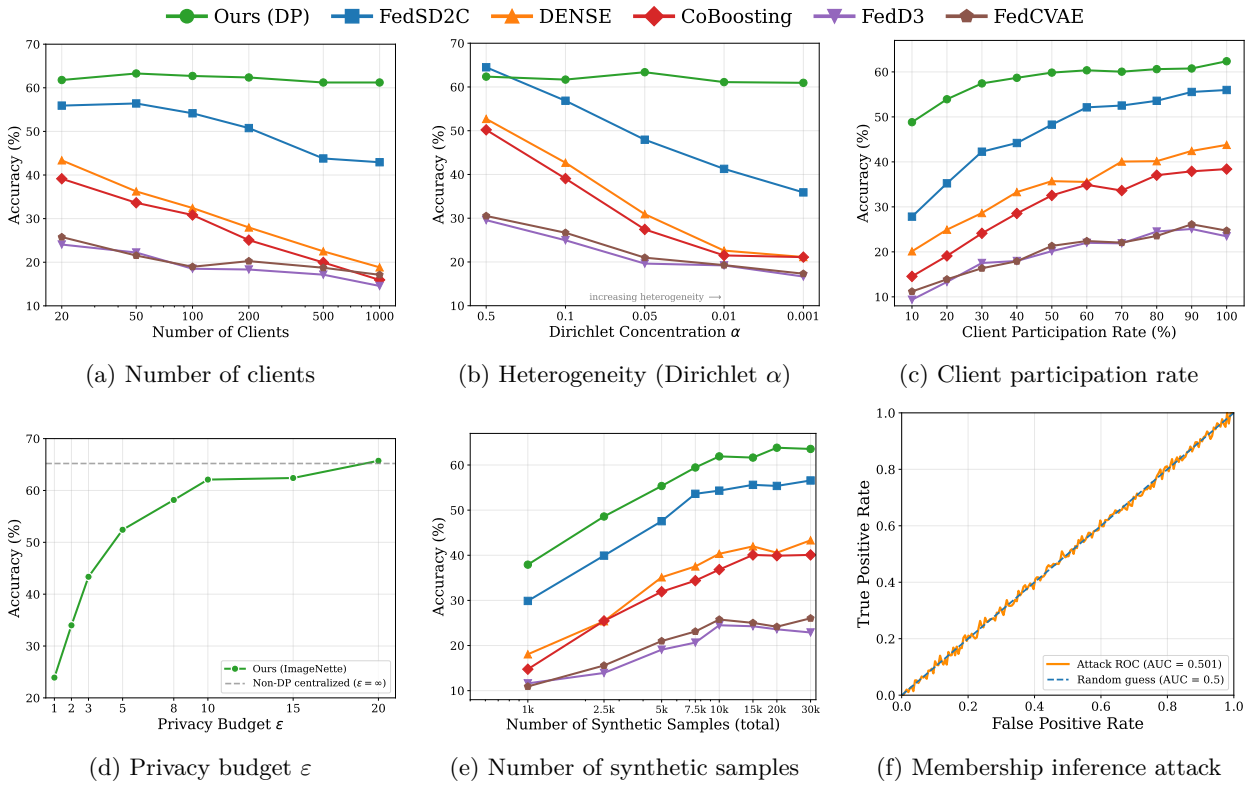
Table 4 shows results on CIFAR-10 and CIFAR-100 under pathological and practical heterogeneity. The value of global synthetic pretraining scales with the severity of label-space fragmentation. On CIFAR-100 pathological, each client observes a narrow class slice and multi-round methods cannot recover structure that was never represented locally. Our globally aggregated statistics reconstruct the full class-conditional distribution, yielding a nearly 9-point lead. This advantage narrows on CIFAR-10 and under practical heterogeneity for the same reason. As local class overlap increases, the information deficit shrinks and local optimization alone suffices.

6 Privacy Evaluation

Membership inference attack. Although the formal DP guarantee already bounds any adversary's advantage, we include a membership inference evaluation Carlini et al. (2022) for comparability with prior OSFL methods that rely on MIA as their primary privacy metric. Figure 3f shows the ROC curve. The AUC is 0.501, indistinguishable from random guessing, which is the optimal result.

This result is naturally follows from the structure of our framework. Unlike standard MIA settings where the attacker probes a model that was directly trained on private data and may have memorized individual examples, our synthetic images are samples from a Gaussian distribution fit to aggregated statistics with DP. To detect whether a specific image participated, an attacker would need its individual contribution to remain detectable after aggregation across hundreds of samples, addition of calibrated DP noise, random sampling from the resulting distribution, and decoding through a public network. Each of these steps dilutes per-record influence, making membership inference structurally difficult.

Privacy-utility tradeoff. Figure 3d shows downstream accuracy on ImageNette as a function of the privacy budget ε (with δ=10^{-5}). At ε=1, heavy noise reduces accuracy to roughly 25%. Performance climbs steeply through ε=3–5 and reaches ∼62% at our default ε=10, with diminishing returns beyond that (the non-DP centralized upper bound is ∼65%). The steep gains at moderate ε reflect the low dimensionality of the latent statistics. Because noise is injected into d=128 dimensional vectors rather than high-dimensional model parameters, even a modest privacy budget yields a favorable signal-to-noise ratio.



(a) Number of clients (b) Heterogeneity (Dirichlet α) (c) Client participation rate

(d) Privacy budget ε (e) Number of synthetic samples (f) Membership inference attack

Figure 3: Ablation studies on ImageNette. (a) Different number of clients. (b) Different heterogeneity settings(c) Participation rate (d) Privacy-utility tradeoff for our method at varying ε. (e) Different number of synthetic samples. (f) MIA ROC curve.

7 Ablation of Method and Experimental Parameters

We ablate key design choices on ImageNette with 20 clients under practical heterogeneity (α=0.1) unless stated otherwise. All baselines are run without DP. Results are shown in Figure 3.

Number of clients. Figure 3a varies the federation size from 20 to 1000. All baselines degrade substantially as data is spread more thinly: FedSD2C drops from 57% to 42%, and weaker methods fall below 20%. Our method is invariant to federation size because the global class-conditional sums are identical regardless of how data is partitioned across clients. The small fluctuations in our curve ($\sim$1%) are purely due to downstream training variance.

Heterogeneity. Figure 3b varies the Dirichlet concentration from α=0.5 (mild heterogeneity) to α=0.001 (extreme). Every baseline degrades monotonically: FedSD2C drops from $\sim$64% to $\sim$36%, and the other methods can drop below 20%. Our method stays at $\sim$62% across all settings because per-class statistics are aggregated identically regardless of how labels are distributed across clients. This heterogeneity invariance is a direct consequence of the additive structure of the pipeline.

Client participation rate. Figure 3c varies the fraction of clients that participate in a single round from 10% to 100%. Our method degrades gracefully. Even at 10% participation (2 out of 20 clients), accuracy remains above 48%, compared to 28% for FedSD2C and below 20% for the other baseline methods. The degradation in our case is small because the statistics from participating clients still combine additively. The main effect of lower participation is fewer samples contributing to the global sums, which increases the relative magnitude of DP noise.

Table 5: **Autoencoder architecture ablation** (ε=10, δ=10^{-5}). d = latent dim; C = per-class sensitivity; $C\sqrt{d}$ = noise-signal proxy.

Codec	Res.	d	C	$C\sqrt{d}$	EuroSAT	CIFAR-10	CIFAR-100	ImageNette	BloodMNIST
TAESD	32	64	7–8	57–68	76.83	70.86	24.70	47.82	72.99
DC-AE f32c32-in	64	128	34.15	387	**80.59**	**73.80**	**28.95**	60.83	**74.73**
DC-AE f64c128-mix	64	128	24.45	277	79.05	73.33	28.48	61.23	74.66
DC-AE f32c32-mix	64	128	21.90	248	79.13	73.53	28.18	59.76	74.70
DC-AE f32c32-in+PCA64	64	64	30.09	241	79.33	70.77	26.13	54.78	72.13
DC-AE f32c32-in+PCA128	128	128	71.07	804	78.39	70.14	27.39	62.38	73.07
DC-AE f32c32-in (raw)	128	512	85.50	1934	58.60	49.14	14.37	38.77	51.67
DC-AE f32c32-in	32	32	20.98	119	75.03	67.11	25.91	53.02	74.34

Number of synthetic samples. Figure 3e varies the total number of synthetic images generated from the DP-protected Gaussians. Accuracy climbs from ∼39% at 1k samples to ∼62% at 10k, then plateaus. The saturation point roughly matches the size of the real dataset (9.5k images), which is expected; once the synthetic dataset is large enough to cover the learned Gaussian distributions, additional samples provide diminishing returns. Baselines follow a similar saturation pattern but at their respectively lower ceilings.

Autoencoder Architecture Ablation Table 5 ablates the autoencoder architecture. The key design question is choosing the latent dimension d. Higher d gives the encoder more capacity to represent images, however the DP noise injected into each coordinate accumulates across dimensions. Specifically, when independent Gaussian noise of scale σ_μ is added to a d-dimensional mean vector, the total L_2 magnitude of the noise increases as $\sigma_\mu\sqrt{d} \propto C\sqrt{d}$, where C is the per-class clipping sensitivity. We report $C\sqrt{d}$ as a signal-to-noise proxy. Lower values mean less total noise relative to the statistics.

The table reveals a clear tradeoff. At d=512 (the raw DC-AE representation at 128×128 resolution), $C\sqrt{d}$=1934 and accuracy collapses to 38–58% across datasets. The latent space is expressive but noise strongly dominates as it overlays the signal. At d=128 with $C\sqrt{d} \approx$ 250–400, accuracy peaks, showing that the latent space is expressive enough and the noise is moderate. Reducing to d=64 via PCA further lowers $C\sqrt{d}$ to 241, but discarding half the latent dimensions sacrifices representational capacity, resulting in 3–7% lower accuracy. At d=32 the noise is minimal ($C\sqrt{d}$=119) but the bottleneck is too narrow to capture class structure.

A second trend is that resolution and reconstruction quality matter independently of d. DC-AE at 64×64 consistently outperforms TAESD at 32×32 despite both using d=64–128, because higher-resolution decoding recovers more visual detail for the downstream classifier. The best overall configuration is DC-AE f32c32-in at 64×64 with d=128, which balances representational capacity, reconstruction quality, and noise tolerance.

8 Limitations

Our method assumes labeled classification tasks where each client knows the class identity of its samples. Extending to unsupervised or self-supervised settings, where no class labels are available to condition the statistics on, remains an open problem for our approach and for data-sharing methods in FL more broadly. Furthermore, the Gaussian model assumes reasonable class separability in the latent space. Under noisy labels or highly overlapping class distributions, the per-class statistics would degrade, though this sensitivity is shared by most class-conditional generative approaches. Moreover, single Gaussian per class could limit the quality of generated data. However, while a GMM would offer more expressive power, it would require per-component assignments that are not additively aggregable across clients without revealing additional information. Finally, classes with very few samples yield noisy statistics. With n_c small, the effective noise in the mean scales as σ_μ/n_c, which can dominate through overlaying the signal. This is, however, inherent to DP mechanisms that calibrate noise to worst-case sensitivity.

9 Conclusion

We have presented FedKT-CSD, a one-shot federated learning framework that generates DP synthetic data by aggregating class-conditional latent statistics from a frozen pretrained autoencoder. The method requires a single communication round, no on-device training, and uploads in the KB regime per client, while providing formal (ε, δ)-DP guarantees with clean accounting via the Gaussian mechanism. Empirically, FedKT-CSD outperforms existing one-shot FL baselines that operate without any privacy protection, is competitive with or superior to centralized DP synthetic data methods, and scales to 1000+ clients with no performance degradation. The synthetic dataset is versatile: we demonstrate its use for both centralized classifier training and downstream personalized FL, where it achieves state-of-the-art results from a single round of communication. The core insight, that pretrained autoencoders provide a shared low-dimensional space in which simple additive statistics suffice for high-quality data generation, suggests a broader design principle for privacy-preserving federated systems.

References

Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pp. 394–403. PMLR, 2018.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkcQFMZRb`.

Mahdi Beitollahi, Alex Bie, Sobhan Hemati, Leo Maxime Brunswic, Xu Li, Xi Chen, and Guojun Zhang. Foundation models meet federated learning: A one-shot feature-sharing method with privacy and performance guarantees. *Transactions on Machine Learning Research*, 2025.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/collins21a.html`.

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

Canh T. Dinh, Nguyen H. Tran, and Tuan D. Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21394–21405, 2020.

Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pp. 1–12, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540359079. doi: 10.1007/11787006_1. URL `https://doi.org/10.1007/11787006_1`.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL `https://doi.org/10.1561/0400000042`.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3557–3568. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/24389bfe4fe2eba8bf9aa9203a44cdad-Paper.pdf.

Chen Gong, Kecen Li, Zinan Lin, and Tianhao Wang. Dpimagebench: A unified benchmark for differentially private image synthesis. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4139–4153, 2025.

Frederik Harder, Kamil Adamczewski, and Mijung Park. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pp. 1819–1827. PMLR, 2021.

Clare Elizabeth Heinbaugh, Emilio Luz-Ricca, and Huajie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_hb4vM3jspB.

Maximilian Andreas Hoefler, Tatsiana Mazouka, Karsten Mueller, and Wojciech Samek. Boosting federated learning with diffusion models for non-iid and imbalanced data. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 7790–7799, 2024. doi: 10.1109/BigData62323.2024.10825355.

Maximilian Andreas Hoefler, Karsten Mueller, and Wojciech Samek. Fedxds: Leveraging model attribution methods to counteract data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4572–4581, 2025.

Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. Pre-text: training language models on private federated data in the age of llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7865–7873, May 2021. doi: 10.1609/aaai.v35i9.16960. URL https://ojs.aaai.org/index.php/AAAI/article/view/16960.

Dihong Jiang, Sun Sun, and Yaoliang Yu. Functional rényi differential privacy for generative modeling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1–2):1–210, June 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Xin-Chun Li, De-Chuan Zhan, Yunfeng Shao, Bingshuai Li, and Shaoming Song. Fedphp: Federated personalization with inherited private models. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 587–602, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86486-6.

Jun Luo and Shandong Wu. Adapt to adaptation: Learning personalization for cross-silo federated learning. In *IJCAI: proceedings of the conference*, volume 2022, pp. 2166, 2022.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: classifer calibration for federated learning with non-iid data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713845393.

H. 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 (AISTATS)*, pp. 1273–1282. PMLR, 2017.

Supriya Nagesh, Justin Y. Chen, Nina Mishra, and Tal Wagner. Private text generation by seeding large language model prompts, 2025. URL `https://arxiv.org/abs/2502.13193`.

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

Benyuan Sun, Hongxing Huo, YI YANG, and Bo Bai. Partialfed: Cross-domain personalized federated learning via partial initialization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23309–23320. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/c429429bf1f2af051f2021dc92a8ebea-Paper.pdf`.

Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4561–4571, 2025.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. Locally differentially private document generation using zero shot prompting. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8442–8457, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.566. URL `https://aclanthology.org/2023.findings-emnlp.566/`.

Yilin Yang, Kamil Adamczewski, Xiaoxiao Li, Danica J Sutherland, and Mijung Park. Differentially private neural tangent kernels (dp-ntk) for privacy-preserving data generation. *Journal of Artificial Intelligence Research*, 81:683–700, 2024.

Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, and Bo Han. Fedfed: Feature distillation against data heterogeneity in federated learning. *Advances in neural information processing systems*, 36:60397–60428, 2023.

Jianqing Zhang, Yang Liu, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Jian Cao. Pfllib: A beginner-friendly and comprehensive personalized federated learning library and benchmark. *Journal of Machine Learning Research*, 26(50):1–10, 2025. URL `http://jmlr.org/papers/v26/23-1634.html`.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. *Advances in Neural Information Processing Systems*, 35: 21414–21428, 2022a.

Junyuan Zhang, Songhua Liu, and Xinchao Wang. One-shot federated learning via synthetic distiller-distillate communication. *Advances in Neural Information Processing Systems*, 37:102611–102633, 2024.

Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10174–10183, June 2022b.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=ehJqJQk9cw`.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zhu21b.html`.