FedHypeVAE: Federated Learning with Hypernetwork-Generated Conditional VAEs for Differentially-Private Embedding Sharing

Sunny Gupta, Nikita Jangid, Amit Sethi
Indian Institute of Technology Bombay, India
{sunnygupta, Nikita Jangid,asethi}@iitb.ac.in

Abstract

Federated learning enables collaborative model development across medical institutions without centralizing sensitive patient data, yet existing embedding-level generative approaches often degrade under non-IID clinical heterogeneity and offer limited formal protection against gradient leakage. We introduce FedHypeVAE, a differentially private, hypernetwork based conditional variational framework that generates client-specific decoders and priors from lightweight, trainable client codes. This bi-level formulation personalizes the generative process while ensuring privacy preserving parameter synthesis decoupled from raw medical images. Federated optimization with differential privacy and distributional alignment strategies improves stability and cross-site generalization. The proposed framework unifies personalization, privacy, and domain adaptability within the generative layer, offering a principled solution for privacy-aware representation learning in multi institutional medical imaging. Code: github.com/sunnyinAI/FedHypeVAE

Introduction

Deep Neural Networks (DNNs) have driven remarkable progress in medical imaging, yet their widespread clinical deployment remains constrained by limited data availability and stringent privacy requirements (Litjens et al. 2017; Shilo, Rossman, and Segal 2020). Medical datasets are often siloed across institutions, while the low prevalence of certain diseases further restricts access to diverse, high-quality training data (Stacke et al. 2020). Although collaborative data sharing could mitigate these challenges, strict regulatory frameworks such as HIPAA and GDPR render centralized dataset aggregation infeasible.

To address these limitations, *Federated Learning* (FL) (McMahan et al. 2017a) has emerged as a distributed paradigm that enables multiple institutions to collaboratively train models without exposing raw data. The classical FedAvg algorithm (McMahan et al. 2017a) aggregates model updates from clients to construct a global model, ensuring that sensitive data remain within institutional boundaries. However, FL faces several persistent challenges. Communication overhead is substantial—especially with high-capacity architectures such as Vision Transformers (ViTs) (Dosovitskiy et al. 2020)—and performance often degrades under non-IID client distributions. Recent efforts to improve efficiency through lightweight architectures (Wu et al. 2023; Xia et al. 2024) have reduced transmission cost but at the expense of robustness and diagnostic fidelity.

An emerging alternative is *synthetic data sharing*, where generative models produce privacy-preserving surrogate datasets instead of transmitting model updates (Koetzier et al. 2024; Ktena et al. 2024). Such methods reduce communication burden and improve cross-domain applicability. While Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and diffusion models (Ho, Jain, and Abbeel 2020) achieve high-fidelity synthesis, they remain unstable or computationally expensive for federated environments. In contrast, Variational Autoencoders (VAEs) and their conditional extensions (CVAEs) offer stable, likelihood-based training and computational efficiency, albeit at the cost of reduced perceptual sharpness. Recent work (Di Salvo et al. 2024) demonstrated that generating data in *embedding space* rather than image space can preserve task-relevant information while mitigating privacy leakage.

This embedding-level paradigm is strengthened by the advent of *foundation encoders* such as DINOv2 (Oquab et al. 2023), which provide compact, semantically rich representations that generalize across imaging domains (Paul and Chen 2022). Training CVAEs on such embeddings enables the generative model to capture diagnostic features efficiently while reducing redundancy and risk of reconstruction-based attacks.

Despite these advances, two fundamental challenges persist. First, existing federated generative frameworks lack the ability to adapt to client-specific heterogeneity, leading to degraded performance under non-IID distributions. Second, formal privacy guarantees are rarely incorporated, with most prior methods relying on heuristic noise injection rather than certified Differential Privacy (DP). Addressing these limitations requires a framework capable of *personalized, differentially-private generative modeling* that remains consistent and generalizable across diverse clinical domains.

To this end, we propose **FedHypeVAE**—a *Federated Hypernetwork-Generated Conditional Variational Autoencoder* designed for privacy-preserving, semantically consistent data synthesis across decentralized medical institu-

tions. Unlike prior embedding-based frameworks that rely on a shared global decoder, FedHypeVAE introduces a unified *hypernetwork* that generates client-specific decoder and class-conditional prior parameters from lightweight private client codes. This design enables client-level personalization while implicitly sharing higher-order generative structure through the hypernetwork, thereby improving adaptability under non-IID conditions. Each client trains a local conditional VAE on embeddings extracted from a frozen foundation model (e.g., DINOv2), while the shared hypernetwork parameters are optimized collaboratively via *Differentially Private Stochastic Gradient Descent* (DP-SGD), ensuring formal (ε, δ)-privacy against gradient inversion and membership inference attacks. Furthermore, a *Maximum Mean Discrepancy (MMD)*-based alignment regularizer enforces cross-site distributional coherence, and a *meta-code synthesis module* learns a domain-agnostic latent code for globally representative embedding generation.

Contributions. Our main contributions are threefold:

- We introduce **FedHypeVAE**, the first federated framework that integrates hypernetwork-based parameter generation with conditional VAEs to enable privacy-preserving embedding synthesis.
- We formulate a principled *bi-level federated optimization* strategy that jointly learns personalized client decoders and a globally consistent hypernetwork under certified (ε, δ)-DP guarantees via gradient clipping and calibrated Gaussian noise.
- We propose an *MMD-based alignment* and *meta-code generation* mechanism that ensure cross-domain coherence and high-fidelity synthetic embedding generation with minimal privacy–utility trade-off.

Extensive multi-institutional experiments on diverse medical imaging datasets demonstrate that **FedHypeVAE** substantially outperforms existing federated generative baselines in terms of robustness, generalization, and privacy compliance. By combining foundation model embeddings, hypernetwork-driven personalization, and differential privacy, FedHypeVAE establishes a new paradigm for secure and effective data sharing in federated medical AI.

Related Work

Gradient inversion and privacy in federated learning

Federated learning (FL) reduces the need for centralized data aggregation by training models through decentralized gradient exchanges across clients. However, a substantial body of research on *gradient inversion* and reconstruction attacks has demonstrated that shared updates (gradients or parameter deltas) can leak sensitive information, including approximate input reconstructions, membership inference, and attribute disclosure (Fredrikson, Jha, and Ristenpart 2015; Zhu, Liu, and Han 2019; Geiping et al. 2020). These risks are amplified in regimes involving highcapacity vision encoders and heterogeneous, small-scale medical datasets, where local gradients become more tightly coupled to individual training samples. This vulnerability motivates defenses that either (i) *minimize the exposure surface* by communicating compressed or less informative representations, or (ii) *alter the communication primitive* so that only aggregated or masked informationrather than raw updatesis revealed to the central server.

Privacy-preserving techniques in federated learning

Privacy-preserving FL methods primarily fall into three methodological categories. **(1) Secure multi-party computation (SMC) and secure aggregation** conceal individual updates by allowing the server to observe only aggregated results, thereby preventing direct reconstruction of any client's gradients (Yao 1982; Bonawitz et al. 2017; Mugunthan et al. 2019; Mou et al. 2021). **(2) Homomorphic encryption (HE)** enables mathematical operations to be performed directly on encrypted parameters, but typically introduces prohibitive computational and communication overhead (Gentry 2009; Park and Lim 2022; Ma et al. 2022). **(3) Differential privacy (DP)** enforces formal privacy guarantees by clipping and perturbing updates with calibrated noise (Geyer, Klein, and Nabi 2017; McMahan et al. 2017b; Yu, Bagdasaryan, and Shmatikov 2020; Bietti et al. 2022; Shen et al. 2023). In addition, empirical defensessuch as gradient pruning, masking, or stochastic noise injection (Zhu, Liu, and Han 2019; Huang et al. 2021; Li et al. 2022; Wei et al. 2020)as well as specialized systems like *Soteria*, *PRECODE*, and *FedKL* (Sun et al. 2020; Scheliga, Mäder, and Seeland 2022; Ren et al. 2023)have been proposed to mitigate leakage. Nonetheless, these techniques often struggle with a persistent *privacyutility trade-off*, where stronger protection degrades model accuracy and cross-domain generalization. Such limitations motivate more structural solutionse.g., hypernetwork-based formulationsthat inherently decouple shared parameters from raw data while maintaining high expressivity (Ha, Dai, and Le 2016).

Federated and Differentially-Private Generative Models

Recent research has explored privacy-preserving data sharing through federated generative modeling. Di Salvo *et al.* (Di Salvo et al. 2024) demonstrated that generating synthetic training data at the *embedding level*, rather than from raw medical images, can preserve data privacy while maintaining high downstream task performance. Building on this principle, the *Embedding-Based Federated Data Sharing via Differentially Private Conditional VAEs* framework (Di Salvo, Nguyen, and Ledig 2025) proposed a federated conditional VAE (CVAE) that learns to synthesize embeddings collaboratively across clients. In their approach, each client trains a CVAE with a symmetric architecturethree linear layers for both the class-conditional encoder and decoderoptimized via a reconstruction loss (mean squared error) and a KullbackLeibler divergence term to regularize the latent distribution towards a standard Gaussian prior. To ensure privacy, differential privacy (DP) noise is added during decoder aggregation using a federated averaging (FedAvg) procedure. This design enables privacy-preserving global generative modeling, yet it relies

on a *shared global decoder*, which can underperform under non-IID data distributions and lacks adaptive capacity across diverse clinical domains.

Hypernetworks for Federated Learning

Hypernetworks have recently gained traction as an effective mechanism for *parameter generation* in federated learning, offering a meta-learning perspective on personalization and model sharing. In this paradigm, a central meta-generator H_ϕ maintained by the server maps a compact client representation e_k to the full parameter set of the client model, $\theta_k = H_\phi(e_k)$ (Ha, Dai, and Le 2016; Shamsian et al. 2021; Carey, Du, and Wu 2022; Li et al. 2023; Tashakori et al. 2023; Lin et al. 2023). This indirect parameterization decouples the global and local learning dynamics: the server learns a global mapping in parameter space, while each client is represented by a low-dimensional embedding capturing its data distribution. As a result, hypernetwork-based federated learning substantially reduces communication and storage overhead, enables smooth interpolation across clients in the embedding space, and provides an elegant mechanism for handling data heterogeneity.

Importantly, this indirection also enhances privacy and robustness. Since the hypernetwork H_ϕ learns a higher-order mapping rather than directly exchanging model gradients, reconstructing raw client data would require jointly inverting both the hypernetwork and the latent client embeddinga substantially harder problem than conventional gradient inversion. Beyond privacy, this architecture offers greater expressivity and adaptability, as the hypernetwork can learn to generate task- or domain-specific parameters that capture client-level inductive biases without explicit parameter sharing. Building on these insights, our proposed **FedHypeVAE** extends the role of hypernetworks beyond discriminative personalization to *generative parameterization*, where H_ϕ produces client-aware decoder and prior parameters for conditional VAEs, thereby enabling privacy-preserving and domain-adaptive data synthesis across heterogeneous medical sites.

Problem Setup and Motivation

We consider a federated system comprising m clients (e.g., medical institutions), indexed by $i \in \{1, \dots, m\}$. Each client privately holds a local embeddinglabel dataset

$$\mathcal{S}_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i},$$

where $x_j^{(i)} \in \mathbb{R}^{d_x}$ denotes a compact feature embedding (typically extracted from a frozen foundation encoder such as DINOv2 (Oquab et al. 2023)) and $y_j^{(i)} \in \mathcal{Y}$ is the corresponding class label. These embeddings serve as a semantically rich, privacy-preserving intermediate representation of raw medical data.

The goal is to collaboratively learn a *federated generative model* that can synthesize globally useful and statistically consistent embeddings across all clients, despite the presence of *non-IID* data heterogeneity. Formally, we aim to approximate the global data distribution $p(x, y)$ through a conditional generative process

$$\hat{x} \sim p_\theta(x \mid z, y), \quad z \sim p_\omega(z \mid y),$$

where (θ, ω) represent the decoder and prior parameters, respectively. In the federated setting, direct sharing of model parameters or data samples is restricted by privacy regulations; hence, each client trains its generative model locally and only communicates privacy-protected information to the central server.

Our proposed **FedHypeVAE** unifies three key components to address this challenge: (i) a *conditional variational autoencoder (CVAE)* that learns the local embedding distribution within each site; (ii) a shared *hypernetwork* H_Φ that maps a lightweight, private client code v_i to client-specific generative parameters (θ_i, ω_i); and (iii) a *federated optimization mechanism* that aggregates knowledge across sites via differentially private stochastic gradient descent (DP-SGD). This formulation enables privacy-preserving personalization within the generative layer while ensuring global coherence and robustness under data heterogeneity.

Methodology

Client-Level Conditional Generative Objective

Each client i models its local embedding distribution $p_i(x|y)$ using a conditional variational autoencoder (CVAE) parameterized by an encoder $q_{\psi_i}(z|x, y)$, a decoder $p_{\theta_i}(x|z, y)$, and a class-conditional prior $p_{\omega_i}(z|y)$. The learning objective maximizes the evidence lower bound (ELBO):

$$\begin{aligned}\mathcal{L}_i^{\text{ELBO}}(\psi_i, \theta_i, \omega_i) = \mathbb{E}_{q_{\psi_i}(z|x,y)}[\log p_{\theta_i}(x|z, y)] \\ - \text{KL}(q_{\psi_i}(z|x, y) \,\|\, p_{\omega_i}(z|y)) \,.\end{aligned} \tag{1}$$

The first term enforces accurate reconstruction of local embeddings, while the KullbackLeibler term regularizes the latent space, promoting smoothness and global consistency across clients. This forms the foundational objective inherited from embedding-based federated CVAE frameworks (Di Salvo, Nguyen, and Ledig 2025; Di Salvo et al. 2024).

Hypernetwork-Based Parameter Generation

To introduce personalization and privacy at the generative layer, we replace independent client decoders with a shared hypernetwork that generates client-specific parameters:

$$\theta_i = h_\theta(v_i; \Phi_\theta), \qquad \omega_i = h_\omega(v_i; \Phi_\omega), \tag{2}$$

where $v_i \in \mathbb{R}^{d_v}$ is a private, trainable client code and $\Phi = \{\Phi_\theta, \Phi_\omega\}$ are shared server-side hypernetwork parameters. This formulation allows each client's generative model to adapt to its domain distribution while decoupling raw data from globally shared parameters, enhancing both privacy and non-IID robustness.

Row-Scaled Efficient Generation. To reduce the parameter footprint, each decoder layer with base weights $W_\ell \in \mathbb{R}^{r_\ell \times c_\ell}$ is modulated by lightweight row-wise scaling and bias shifting:

$$\begin{aligned}W_\ell(v_i) &= \text{diag}(d_\ell(v_i))\, W_\ell, \\ b_\ell(v_i) &= b_\ell + \Delta b_\ell(v_i),\end{aligned} \tag{3}$$

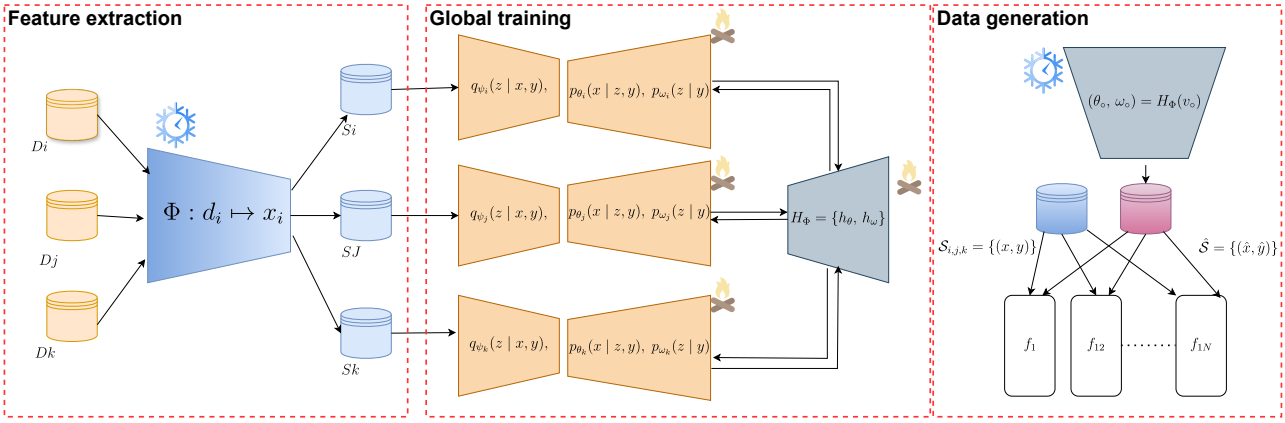

Figure 1: Overview of the proposed **FedHypeVAE** framework. (1) Each participating client $\mathcal{H}_i$ transforms its local image dataset $\mathcal{D}_i$ into an embedding-level dataset $\mathcal{S}_i$ using a frozen foundation encoder Φ, substantially reducing communication and storage cost. (2) Locally, each client trains a conditional variational autoencoder (CVAE) parameterized by an encoder–decoder pair $(q_{\psi_i}, p_{\theta_i})$ and a class-conditional prior p_{ω_i}, which model the embedding distribution without exposing raw data. (3) A server-side hypernetwork $H_\Phi = \{h_\theta, h_\omega\}$ maps private client codes v_i to client-specific decoder and prior parameters, and is optimized federatively via differentially-private stochastic gradient descent (DP-SGD). (4) After convergence, a neutral meta-code $v_\circ$ produces a global decoder–prior pair $(\theta_\circ, \omega_\circ)$ that generates synthetic embeddings $\hat{\mathcal{S}} = \{(\hat{x}, \hat{y})\}$, which can be combined with local data for downstream models $f_1, \ldots, f_N$.

where $d_\ell(v_i)$ and $\Delta b_\ell(v_i)$ are predicted by h_θ. This strategy follows the HyperLSTM principle (Ha, Dai, and Le 2016), retaining expressivity while minimizing computation and communication overhead.

Hyper-Generated Class Priors. Similarly, the class-conditional Gaussian priors are generated as

$$\begin{aligned}(\mu_{i,y}, \log \sigma_{i,y}) &= g_\omega(h_\omega(v_i; \Phi_\omega), e(y)),\\ p_{\omega_i}(z|y) &= \mathcal{N}(\mu_{i,y}, \mathrm{diag}(\sigma^2_{i,y})),\end{aligned} \tag{4}$$

where $e(y)$ is a learnable label embedding. This parameterization enables the model to capture domain-specific feature styles and better calibrate latent priors across sites.

Stability Regularization and Cross-Site Alignment

Each client minimizes a stability-regularized objective that combines the negative ELBO with structural constraints:

$$\begin{aligned}\mathcal{J}_i(\psi_i, v_i; \Phi) = -\mathbb{E}_{(x,y)\sim\mathcal{S}_i}\left[\mathcal{L}_i^{\text{ELBO}}\right]\\ + \lambda_{\text{Lip}}\mathcal{R}_{\text{Lip}}(h_\theta, h_\omega) + \lambda_v\|v_i\|_2^2,\end{aligned} \tag{5}$$

where $\mathcal{R}_{\text{Lip}}$ enforces spectral-norm or Jacobian control for Lipschitz stability, and λ_v constrains client code magnitudes.

Cross-Site Distribution Alignment. To align real and synthetic embeddings, each client computes a local Maximum Mean Discrepancy (MMD) loss:

$$\begin{aligned}\text{MMD}_i^2 = \frac{1}{|\mathcal{X}_i|^2}\sum_{x,x'\in\mathcal{X}_i} k(x, x') + \frac{1}{|\hat{\mathcal{X}}_i|^2}\sum_{\hat{x},\hat{x}'\in\hat{\mathcal{X}}_i} k(\hat{x}, \hat{x}')\\ - \frac{2}{|\mathcal{X}_i||\hat{\mathcal{X}}_i|}\sum_{x\in\mathcal{X}_i,\hat{x}\in\hat{\mathcal{X}}_i} k(x, \hat{x}),\end{aligned} \tag{6}$$

where $k(\cdot, \cdot)$ is a Gaussian multi-kernel function. This term promotes consistent latent distributions across domains without requiring any raw data exchange.

Federated Hypernetwork Optimization under Differential Privacy

The shared hypernetwork parameters Φ are optimized collaboratively across clients via DP-SGD. The global federated objective aggregates client losses and alignment regularizers:

$$\min_\Phi \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{(x,y)\sim\mathcal{S}_i}[\mathcal{J}_i(\psi_i^\star, v_i^\star; \Phi)] + \lambda_{\text{MMD}}\mathbb{E}[\text{MMD}_i^2]. \tag{7}$$

Each client optimizes its local encoder and code parameters:

$$\begin{aligned}\psi_i &\leftarrow \psi_i - \eta_\psi\nabla_{\psi_i}(-\mathcal{L}_i^{\text{ELBO}}),\\ v_i &\leftarrow v_i - \eta_v\nabla_{v_i}(-\mathcal{L}_i^{\text{ELBO}} + \lambda_v\|v_i\|_2^2).\end{aligned} \tag{8}$$

Differentially Private Gradient Construction. For each minibatch B_i, client i computes a per-sample gradient, clips it to a bound C, and adds Gaussian noise:

$$\tilde{g}_i = \frac{1}{|B_i|}\sum_{(x,y)\in B_i}\text{clip}(\nabla_\Phi\mathcal{J}_i, C) + \mathcal{N}(0, \sigma^2C^2I). \tag{9}$$

Only these noise-perturbed gradients $\tilde{g}_i$ are sent to the server, ensuring (ε, δ)-differential privacy while keeping ψ_i, v_i, and raw data local.

Server-Side Aggregation. The server aggregates privatized gradients in a FedAvg-style update:

$$\Phi \leftarrow \Phi - \eta_\Phi\sum_{i=1}^m w_i\tilde{g}_i, \quad w_i = \frac{n_i}{\sum_j n_j}. \tag{10}$$

Table 1: **Comparison of federated baselines and our proposed FedHypeVAE across Abdominal CT and ISIC 2025 datasets.** Values denote mean ± standard deviation of Accuracy (ACC) and Balanced Accuracy (BACC) across clients over three seeds. The best performance for each dataset configuration is shown in **bold**.

Method	CT (IID)		CT (α = 0.3)		ISIC 2025 (IID)		ISIC 2025 (α = 0.3)	
	ACC (%)	BACC (%)	ACC (%)	BACC (%)	ACC (%)	BACC (%)	ACC (%)	BACC (%)
FedAvg	73.27±1.18	67.04±1.21	64.91±5.83	58.68±2.96	61.20±2.8	54.10±2.5	61.00±2.8	54.00±2.5
FedProx	73.30±1.16	66.88±1.20	64.81±6.18	58.61±5.80	61.75±3.0	54.60±2.8	60.90±3.0	53.90±2.8
FedLambda	77.27±0.83	71.26±0.87	81.10±3.76	59.02±2.59	63.30±2.7	55.20±2.8	76.25±3.2	54.54±2.6
DP-CGAN	77.54±1.42	71.99±1.14	88.91±2.04	57.44±2.46	64.80±2.9	55.80±3.0	83.12±2.9	53.38±2.9
DP-CVAE (paper)	77.60±0.72	71.77±0.85	88.88±1.41	57.63±3.29	66.20±2.8	56.30±2.6	83.10±2.8	53.46±2.7
FedHypeVAE (ours)	**81.32**±1.05	**76.08**±1.12	**90.09**±1.07	**62.14**±1.02	**67.70**±2.6	**56.90**±2.7	**84.00**±2.6	**57.74**±2.8

This completes one communication round under formal DP guarantees.

Global Meta-Code Synthesis and Generation

After convergence, the server learns a *neutral meta-code* $v_\circ$ using DP-noisy global statistics $\{\hat{\mu}_y, \hat{\Sigma}_y\}$:

$$v_\circ = \arg\min_v \sum_{y\in\mathcal{Y}} \|\mathbb{E}_{z\sim p_{\omega_\circ}(z|y)}[x(z,y)] - \hat{\mu}_y\|_2^2 + \beta\|\mathrm{Cov}_z[x(z,y)] - \hat{\Sigma}_y\|_F^2. \tag{11}$$

Synthetic embeddings are generated as

$$\hat{x} \sim p_{\theta_\circ}(x|z,y), \quad \theta_\circ = h_\theta(v_\circ;\Phi), \quad \omega_\circ = h_\omega(v_\circ;\Phi), \tag{12}$$

where $z \sim \mathcal{N}(0, I)$. This meta-code enables controllable, domain-agnostic synthesis under privacy constraints.

Mixture of Meta-Codes. For richer global synthesis, K meta-codes $\{v_k\}_{k=1}^K$ with mixture weights $\pi_k \geq 0$, $\sum_k \pi_k = 1$ can be used:

$$\theta_{\text{mix}} = \sum_{k=1}^K \pi_k h_\theta(v_k;\Phi), \quad \omega_{\text{mix}} = \sum_{k=1}^K \pi_k h_\omega(v_k;\Phi), \quad \hat{x} \sim p_{\theta_{\text{mix}}}(x|z,y). \tag{13}$$

We impose spectral-norm constraints on (h_θ, h_ω) for Lipschitz stability, bound $\|v_i\|_2 \leq r$, and track privacy loss using moments accounting over (q, σ, T). Cross-site MMD alignment mitigates non-IID drift, while mixture meta-codes improve global coverage and diversity.

Algorithm 1: **FedHypeVAE**

Require: Number of clients m; privacy budget (ε, δ); learning rates $\eta_\psi, \eta_v, \eta_\Phi$; clipping bound C; noise scale σ; regularization weights $\lambda_{\text{MMD}}, \lambda_{\text{Lip}}, \lambda_v$

Initialize shared hypernetwork parameters $\Phi = \{\Phi_\theta, \Phi_\omega\}$, local encoders ψ_i, and private codes $v_i \sim \mathcal{N}(0, I)$ for each client i.
for each communication round $t = 1$ to T **do**
  Client-side (for each i in parallel):
  Sample minibatch $B_i \subseteq \mathcal{S}_i$.
  Compute local ELBO loss $\mathcal{L}_i^{\text{ELBO}}$ (Eq. 1).
  Update local encoder ψ_i and client code v_i (Eq. 9).
  Evaluate alignment loss MMD_i^2 (Eq. 6).
  Compute privatized gradient:

$$\tilde{g}_i = \frac{1}{|B_i|} \sum_{(x,y)\in B_i} \text{clip}(\nabla_\Phi \mathcal{J}_i, C) + \mathcal{N}(0, \sigma^2 C^2 I).$$

  Transmit $\tilde{g}_i$ to the server.
  Server-side:
  Aggregate and update global hypernetwork:

$$\Phi \leftarrow \Phi - \eta_\Phi \sum_{i=1}^m w_i \tilde{g}_i, \quad w_i = \frac{n_i}{\sum_j n_j}.$$

end for
Post-training: Learn meta-code $v_\circ$ (Eq. 12); generate synthetic embeddings $\hat{x} \sim p_{\theta_\circ}(x|z,y)$ where $\theta_\circ = h_\theta(v_\circ;\Phi)$ and $\omega_\circ = h_\omega(v_\circ;\Phi)$; optionally mix K meta-codes (Eq. 13).

Ensure: Trained global hypernetwork Φ and synthetic dataset $\hat{\mathcal{S}}$.

Experimental results

Experimental Settings

Datasets and Metrics. We evaluate FedHypeVAE on two complementary multi-site medical imaging benchmarks. (1) The ISIC 2025 MILK10k dataset (Philipp et al. 2025) comprises 10,000 dermoscopic images annotated across multiple diagnostic categories, simulating a multi-institutional skin-lesion federation. (2) The Abdominal CT (Sagittal view) dataset (Xu et al. 2019) contains 25,211 CT slices across 11 anatomical classes and is widely adopted in cross-organ localization tasks. Following recent FL studies (Li et al. 2024; Chen et al. 2023), each dataset is distributed among $m = 10$ clients under both IID and heterogeneous settings using a Dirichlet partition with $\alpha = 0.3$. Raw medical images are converted into compact feature embeddings $\mathcal{S}_i = \{(x, y)\}$ using a frozen DINOv2 encoder (Oquab et al. 2023), ensuring representation consistency while preserving privacy. Evaluation metrics include per-client *accuracy* and *balanced accuracy (BACC)*, averaged over three random seeds to assess robustness under domain skew.

Implementation Details. All downstream classifiers are implemented as single-layer linear models on top of DI-

NOv2 embeddings (Oquab et al. 2023). FedHypeVAE and all baselines are trained for 50 communication rounds with 5 local epochs per round using SGD ($\eta = 10^{-3}$). Differential privacy is enforced via DP-SGD using the OPACUS library (Yousefpour et al. 2021) with $(\varepsilon, \delta) = (1.0, 10^{-4})$ and clipping norm 1.5, providing formal privacy guarantees (Nasr et al. 2021; Lange et al. 2022). Comparative baselines include FedAvg and **FedProx** (Li et al. 2020), alongside a DP-CVAE variant (Di Salvo et al. 2024). All models are trained and evaluated under identical federation settings.

Results and Discussion

Table 1 reports results across both datasets under IID and non-IID conditions. **FedHypeVAE** consistently surpasses baseline federated classifiers in terms of generative fidelity, accuracy, and balanced accuracy. Its hypernetwork-based decoder and prior generation enable client-adaptive modeling, while the MMD alignment term mitigates cross-site distribution drift. Even under strict privacy budgets ($\varepsilon \leq 3.0$, $\delta = 10^{-5}$), the model preserves high reconstruction fidelity and generalization, outperforming DP-CVAE in both radiological and dermatological domains. Unlike parameter-regularization-based personalization methods (Marfoq et al. 2022), FedHypeVAE achieves personalization directly within the generative layer, producing semantically consistent, privacy-preserving embeddings across diverse modalities.

Results and Discussion

FedHypeVAE was evaluated on multi-site medical imaging datasets under both IID and non-IID partitions, showing consistent gains in generative fidelity, robustness, and privacy over federated CVAE baselines (Di Salvo, Nguyen, and Ledig 2025; Di Salvo et al. 2024; Pfitzner and Arnrich 2022). Under comparable privacy budgets ($\varepsilon \leq 3.0$, $\delta = 10^{-5}$), it achieves higher accuracy and balanced accuracy while preserving strict differential privacy guarantees. These improvements stem from the hypernetwork's ability to generate client-adaptive decoder and prior parameters that capture local variations without degrading global coherence. The inclusion of *MMD-based cross-site alignment* stabilizes latent representations across heterogeneous domains, mitigating embedding drift typical in federated settings. Moreover, gradient-level *DP-SGD* ensures a superior privacy–utility trade-off compared to weight-level noise injection, maintaining reconstruction quality under strong privacy constraints. Collectively, **FedHypeVAE** advances differentially private generative learning by achieving domain-consistent, semantically faithful, and privacy-compliant embedding synthesis across decentralized medical datasets.

Conclusion

We presented **FedHypeVAE**, a hypernetwork-driven, bi-level federated generative framework that extends embedding-based differentially-private CVAE paradigms toward adaptive, privacy-preserving data synthesis. By introducing a shared hypernetwork that generates client-specific decoder and prior parameters from lightweight private codes, FedHypeVAE achieves fine-grained personalization without compromising data confidentiality. The incorporation of cross-site MMD alignment and meta-code synthesis ensures coherent global representation under severe non-IID conditions, while DP-SGD guarantees formal (ε, δ)-privacy throughout training. Collectively, these advances establish a unified approach that bridges generative modeling, personalization, and differential privacysetting a foundation for secure, generalizable, and data-efficient collaboration across medical institutions.

References

Bietti, A.; Wei, C.-Y.; Dudik, M.; Langford, J.; and Wu, S. 2022. Personalization improves privacy-accuracy tradeoffs in federated learning. In *International Conference on Machine Learning*, 1945–1962. PMLR.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

Carey, A. N.; Du, W.; and Wu, X. 2022. Robust personalized federated learning under demographic fairness heterogeneity. In *2022 IEEE International Conference on Big Data (Big Data)*, 1425–1434. IEEE.

Chen, M.; Jiang, M.; Dou, Q.; Wang, Z.; and Li, X. 2023. Fedsoup: Improving generalization and personalization in federated learning via selective model interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 318–328. Springer.

Di Salvo, F.; Nguyen, H. H. M.; and Ledig, C. 2025. Embedding-Based Federated Data Sharing via Differentially Private Conditional VAEs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 138–147. Springer.

Di Salvo, F.; Tafler, D.; Doerrich, S.; and Ledig, C. 2024. Privacy-preserving datasets by capturing feature distributions with Conditional VAEs. *arXiv preprint arXiv:2408.00639*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in neural information processing systems*, 33: 16937–16947.

Gentry, C. 2009. *A fully homomorphic encryption scheme*. Stanford university.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Ha, D.; Dai, A.; and Le, Q. V. 2016. Hypernetworks. *arXiv preprint arXiv:1609.09106*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, Y.; Gupta, S.; Song, Z.; Li, K.; and Arora, S. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in neural information processing systems*, 34: 7232–7241.

Koetzier, L. R.; Wu, J.; Mastrodicasa, D.; Lutz, A.; Chung, M.; Koszek, W. A.; Pratap, J.; Chaudhari, A. S.; Rajpurkar, P.; Lungren, M. P.; et al. 2024. Generating synthetic data for medical imaging. *Radiology*, 312(3): e232471.

Ktena, I.; Wiles, O.; Albuquerque, I.; Rebuffi, S.-A.; Tanno, R.; Roy, A. G.; Azizi, S.; Belgrave, D.; Kohli, P.; Cemgil, T.; et al. 2024. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4): 1166–1173.

Lange, L.; Schneider, M.; Christen, P.; and Rahm, E. 2022. Privacy in Practice: Private COVID-19 Detection in X-Ray Images (Extended Version). *arXiv preprint arXiv:2211.11434*.

Li, H.; Cai, Z.; Wang, J.; Tang, J.; Ding, W.; Lin, C.-T.; and Shi, Y. 2023. FedTP: Federated learning by transformer personalization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13426–13440.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

Li, X.; Zhang, W.; Yu, Y.; Zheng, W.-S.; Zhang, T.; and Wang, R. 2024. SiFT: A Serial Framework with Textual Guidance for Federated Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 655–665. Springer.

Li, Z.; Zhang, J.; Liu, L.; and Liu, J. 2022. Auditing privacy defenses in federated learning via generative gradient leakage. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10132–10142.

Lin, Y.; Wang, H.; Li, W.; and Shen, J. 2023. Federated learning with hyper-network—A case study on whole slide image analysis. *Scientific Reports*, 13(1): 1724.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.

Ma, J.; Naas, S.-A.; Sigg, S.; and Lyu, X. 2022. Privacy-preserving federated learning based on multi-key homomorphic encryption. *International Journal of Intelligent Systems*, 37(9): 5880–5901.

Marfoq, O.; Neglia, G.; Vidal, R.; and Kameni, L. 2022. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, 15070–15092. PMLR.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017a. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

McMahan, H. B.; Ramage, D.; Talwar, K.; and Zhang, L. 2017b. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Mou, W.; Fu, C.; Lei, Y.; and Hu, C. 2021. A verifiable federated learning scheme based on secure multi-party computation. In *International conference on wireless algorithms, systems, and applications*, 198–209. Springer.

Mugunthan, V.; Polychroniadou, A.; Byrd, D.; and Balch, T. H. 2019. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA.

Nasr, M.; Songi, S.; Thakurta, A.; Papernot, N.; and Carlin, N. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, 866–882. IEEE.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, J.; and Lim, H. 2022. Privacy-preserving federated learning using homomorphic encryption. *Applied Sciences*, 12(2): 734.

Paul, S.; and Chen, P.-Y. 2022. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, 2071–2081.

Pfitzner, B.; and Arnrich, B. 2022. Dpd-fvae: Synthetic data generation using federated variational autoencoders with differentially-private decoder. *arXiv preprint arXiv:2211.11591*.

Philipp, T.; Bengü, A. N.; Cliff, R.; Veronica, R.; Verche, T.; Jochen, W.; Katharina, W. A.; Christoph, M.; Nicholas, K.; Allan, H.; et al. 2025. MILK10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. *Journal of Investigative Dermatology*.

Ren, H.; Deng, J.; Xie, X.; Ma, X.; and Ma, J. 2023. Gradient leakage defense with key-lock module for federated learning. *arXiv preprint arXiv:2305.04095*.

Scheliga, D.; Mäder, P.; and Seeland, M. 2022. Precode – a generic model extension to prevent deep gradient leakage. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1849–1858.

Shamsian, A.; Navon, A.; Fetaya, E.; and Chechik, G. 2021. Personalized federated learning using hypernetworks. In *International conference on machine learning*, 9489–9502. PMLR.

Shen, Z.; Ye, J.; Kang, A.; Hassani, H.; and Shokri, R. 2023. Share your representation only: Guaranteed improvement of the privacy-utility tradeoff in federated learning. *arXiv preprint arXiv:2309.05505*.

Shilo, S.; Rossman, H.; and Segal, E. 2020. Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine*, 26(1): 29–38.

Stacke, K.; Eilertsen, G.; Unger, J.; and Lundström, C. 2020. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics*, 25(2): 325–336.

Sun, J.; Li, A.; Wang, B.; Yang, H.; Li, H.; and Chen, Y. 2020. Provable defense against privacy leakage in federated learning from representation perspective. *arXiv preprint arXiv:2012.06043*.

Tashakori, A.; Zhang, W.; Wang, Z. J.; and Servati, P. 2023. SemiPFL: Personalized semi-supervised federated learning framework for edge intelligence. *IEEE Internet of Things Journal*, 10(10): 9161–9176.

Wei, W.; Liu, L.; Loper, M.; Chow, K.-H.; Gursoy, M. E.; Truex, S.; and Wu, Y. 2020. A framework for evaluating gradient leakage attacks in federated learning. *arXiv preprint arXiv:2004.10397*.

Wu, N.; Yu, L.; Yang, X.; Cheng, K.-T.; and Yan, Z. 2023. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 692–702. Springer.

Xia, Y.; Ma, B.; Dou, Q.; and Xia, Y. 2024. Enhancing federated learning performance fairness via collaboration graph-based reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263–272. Springer.

Xu, X.; Zhou, F.; Liu, B.; Fu, D.; and Bai, X. 2019. Efficient multiple organ localization in CT image using 3D region proposal network. *IEEE transactions on medical imaging*, 38(8): 1885–1898.

Yao, A. C. 1982. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, 160–164. IEEE.

Yousefpour, A.; Shilov, I.; Sablayrolles, A.; Testuggine, D.; Prasad, K.; Malek, M.; Nguyen, J.; Ghosh, S.; Bharadwaj, A.; Zhao, J.; et al. 2021. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*.

Yu, T.; Bagdasaryan, E.; and Shmatikov, V. 2020. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.