FissionVAE: Federated Non-IID Image Generation with Latent Space and Decoder Decomposition

Chen Hu[1] , **Hanchi Ren**[1] , **Jingjing Deng**[2] , **Xianghua Xie**[1,*] and **Xiaoke Ma**[3]
[1]Swansea University, United Kingdom
[2]Durham University, United Kingdom
[3]Xi'Dian University, P. R. China
{chen.hu, hanchi.ren, x.xie}@swansea.ac.uk, jingjing.deng@durham.ac.uk

Abstract

Federated learning is a machine learning paradigm that enables decentralized clients to collaboratively learn a shared model while keeping all the training data local. While considerable research has focused on federated image generation, particularly Generative Adversarial Networks, Variational Autoencoders have received less attention. In this paper, we address the challenges of non-IID (independently and identically distributed) data environments featuring multiple groups of images of different types. Non-IID data distributions can lead to difficulties in maintaining a consistent latent space and can also result in local generators with disparate texture features being blended during aggregation. We thereby introduce FissionVAE that decouples the latent space and constructs decoder branches tailored to individual client groups. This method allows for customized learning that aligns with the unique data distributions of each group. Additionally, we incorporate hierarchical VAEs and demonstrate the use of heterogeneous decoder architectures within FissionVAE. We also explore strategies for setting the latent prior distributions to enhance the decoupling process. To evaluate our approach, we assemble two composite datasets: the first combines MNIST and FashionMNIST; the second comprises RGB datasets of cartoon and human faces, wild animals, marine vessels, and remote sensing images. Our experiments demonstrate that FissionVAE greatly improves generation quality on these datasets compared to baseline federated VAE models.

1 Introduction

Generative models have attracted increasing attention in recent years due to their impressive ability to generate new data across various modalities, including images [Ho *et al.*, 2020], texts [Touvron *et al.*, 2023], and audios [Borsos *et al.*, 2023]. As these models, like other deep learning systems, require substantial amounts of data, concerns regarding data privacy have elevated among regulatory authorities and the public. Unlike the traditional centralized learning paradigm, which collects all data on a single computer system for training, federated learning allows private data to remain on the owner's device. In this paradigm, local devices train models independently, and a central server aggregates these models without accessing the individual data directly. Although this distributed approach enhances privacy protection, it also introduces unique challenges not encountered in centralized systems. Since data remains distributed across various client devices, the training samples are not guaranteed to be identically distributed. This can lead to inconsistencies in learning objectives among clients, resulting in degraded performance when these models are aggregated on the server.

In the context of FL with non-IID data, generative models such as Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2014] and Variational Autoencoders (VAEs) [Kingma and Welling, 2014] face additional challenges. These models involve sampling from a latent distribution, and the generator or decoder trained on client devices may develop differing interpretations of the same latent space. This discrepancy can lead to difficulties in maintaining a consistent and unified latent space, resulting in ambiguous latent representations. A further challenge arises from the role of the generator or decoder, which are tasked with mapping latent inputs to the sample space by synthesizing the shape, texture, and colors of images. Aggregating generative models trained on non-IID image data can produce artifacts that appear as a blend of disparate image types, because generators trained on non-IID local data capture the characteristics of varied visual features. Specifically for GANs, another problem arises from local discriminators, which may provide conflicting feedback that hinders model convergence. With the limited data available in FL settings, discriminators can quickly overfit to the training samples [Karras *et al.*, 2020]. If an updated generator from the server produces images of classes not present in a client's local dataset, the local discriminator might incorrectly label well-generated images as fake, simply because they do not match the local data distribution. This mislabeling can significantly impede the generator's ability to synthesize realistic images.

Existing research on generative models for non-IID data in federated learning (FL) has primarily focused on GANs. MDGAN [Hardy *et al.*, 2019] proposes exchanging local dis-

* Corresponding author.

criminators among clients during training. This strategy allows discriminators to access a broader spectrum of local data, thereby avoiding biased feedback to the generator. The authors of [Yonetani *et al.*, 2019] uses the local discriminator that gives the highest score to a generated sample to update the global generator, promoting the idea that local discriminators should only judge samples from familiar distributions. In [Xiong *et al.*, 2023], the authors aggregate generators at the group level for client groups sharing similar data distributions before performing a global aggregation, then the global generator is aggregated similar to [Yonetani *et al.*, 2019]. Both [Yonetani *et al.*, 2019] and [Xiong *et al.*, 2023] involve sending synthesized samples back to local clients, which could potentially increase the risk of compromising client data privacy.

Studies employing VAEs solely for image generation purposes are less common. The works in [Chen and Vikalo, 2023] and [Heinbaugh *et al.*, 2023] utilize VAEs to produce synthetic images that assist in training global classifiers. In [Chen and Vikalo, 2023], the global decoder generates minority samples for local classifiers by sampling from class means with added noise. The approach in [Heinbaugh *et al.*, 2023] treats converged local decoders as teacher models and uses knowledge distillation to train a global generator on the server side without further local updates. While this decoder can produce useful samples for classification tasks, it risks overfitting to the potentially flawed output from local decoders and lacks generative diversity, which is crucial for high-quality image generation. Recent studies [Bohacek and Farid, 2023] [Shumailov *et al.*, 2024] have shown that generative models trained on generated samples instead of real data are prone to collapsing. VAEs are also widely used in collaborative filtering tasks for recommendation systems [Polato, 2021; Zhang *et al.*, 2024; Li *et al.*, 2025]. These models typically learn user embeddings from interaction vectors using a standard Gaussian prior, and decode into item-score distributions for ranking. In contrast, image generation tasks require decoding into high-dimensional pixel space, where issues such as latent space ambiguity and domain-specific texture blending and arise, which are not present in collaborative filtering. As such, the architectural and modeling considerations in our work are fundamentally different.

In response to the challenges posed by non-IID data in federated image generation, we introduce a model named FissionVAE. This model is specifically tailored to environments featuring multiple groups of images of different types. To mitigate the problem of mixed latent space interpretation, FissionVAE decomposes the latent space into distinctive priors, hence adapting to the diverse data distributions across different image types. We further refine this approach by investigating strategies for encoding the prior Gaussians. Additionally, to prevent the blending of unrelated visual features in the generated outputs, FissionVAE employs specialized decoder branches for each client group. This method not only accommodates the unique characteristics of each data subset but also enhances the model's generative capabilities in highly heterogeneous environments. The primary contributions of our research are detailed as follows:

1. We introduce FissionVAE for federated non-IID image generation. In FissionVAE, we decompose the latent space according to the distinct data distributions of client groups. This approach ensures that each client's data are mapped to its corresponding latent distribution without the adverse effects of averaging dissimilar distributions during aggregation. Moreover, by implementing separate decoder branches for different groups of data, FissionVAE allows for specialized generation tailored to different image types, which is crucial for preserving the distinct visual features of different image types during the generative process.

2. We explore various strategies for encoding Gaussian priors to enhance the effectiveness of latent space decomposition. We further extends FissionVAE by introducing the hierarchical inference architecture. We demonstrate that with the decomposed decoder branches, it is feasible to employ heterogeneous decoder architectures in FissionVAE, allowing for more flexible model deployment on clients.

3. We validate FissionVAE with extensive experiments on two composite datasets combining MNIST with FashionMNIST, and a more diverse set comprising cartoon and human faces, animals, marine vessels, and remote sensing images. Our results demonstrate improvements in generation quality over the existing baseline federated VAE.

The remainder of the paper is organized as follows: In Section 2, we describe the baseline FedVAE model and the FissionVAE variants we propose. Section 3 presents the experimental setup, including the configuration details and an analysis of the results. Finally, we conclude the paper in Section 4 with a summary of our findings and a discussion on potential future directions.

2 Investigating Strategies for Non-IID Image Generation with VAEs

In this section, we describe our methodology for exploring VAE configurations tailored for generating images under non-IID conditions in a federated learning framework. For background on FL and VAEs, please refer to the supplementary material. We specifically address scenarios where clients are categorized based on distinct data distributions. For illustrative purposes, we consider the case where some clients exclusively possess hand-written digit images from the MNIST dataset, while others maintain only clothing images from the FashionMNIST dataset. We follow to the standard federated learning framework, wherein a central server is tasked with aggregating updates from the clients and subsequently distributing the updated model back to them. FedAvg [McMahan *et al.*, 2023] is employed for server-side aggregation. Each client retains a subset of data representative of its respective group and conducts local training independently. A more practical scenario with RGB images and a larger number of client groups is explored and discussed in the experiments section (Section 3).

2.1 FedVAE

A straightforward strategy for implementing VAEs in federated learning is using a unified encoder-decoder architecture. In this configuration, all clients share a common latent space (often predefined as the normal distribution $\mathcal{N}(0, 1)$) and the

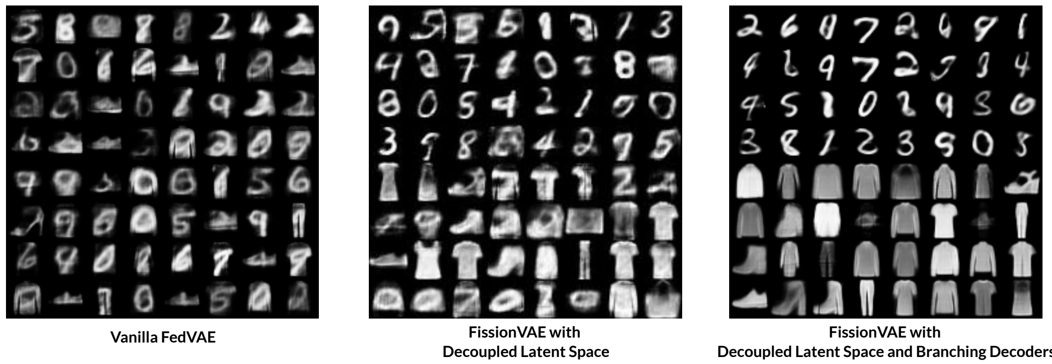

Figure 1: Qualitative results of the baseline FedVAE and proposed FissionVAEs. As we further decoupling the latent space and decoders in the federated environment, the quality of generated images is improved.

central server indiscriminately aggregates client models at the end of each training round. This approach is named FedVAE in [Jiang *et al.*, 2023] for trajectory data generation. Fig. 2 illustrates this baseline training scheme.

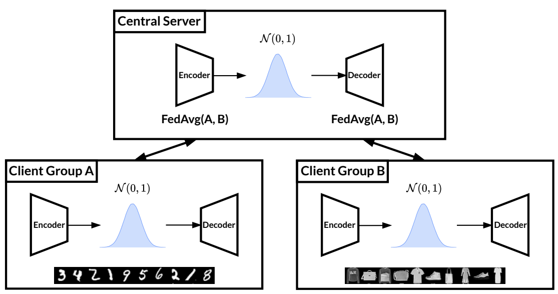

Figure 2: An illustration of baseline FedVAE. The encoder and the decoder of the VAE are aggregated through FedAvg regardless of their client groups.

Despite the simplicity of this strategy, it present significant challenges in the non-IID scenario. Specifically, employing a single prior distribution for the latent space does not account for the distinct data distributions across different clients. Encoders from different client groups may map their uniquely distributed data into the same region of the latent space. Consequently, client decoders might interpret this shared latent space differently, leading to inconsistencies or even conflicts among client models during aggregation at the server. Figure 1 shows randomly generated samples produced after training the federated Vanilla VAE on the combined dataset of MNIST and FashionMNIST. These samples clearly exhibit artifacts that appear to blend features of handwritten digits with clothing items, indicating the aggregation conflicts inherent in this method.

2.2 FissionVAE with Latent Space Decoupling

To address the conflicting latent space issue identified above, we propose decomposing the latent space according to different data groups, while maintaining a unified architecture for the encoder and decoder. This approach corresponds to the architecture shown in Fig. 3.

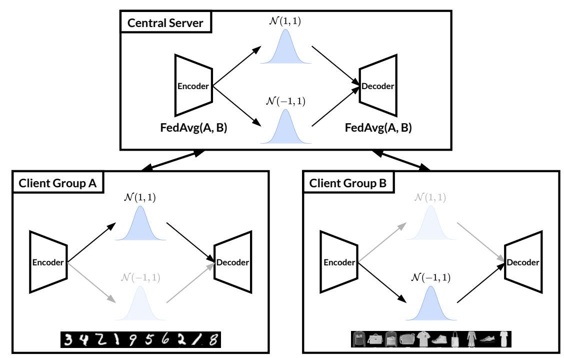

Figure 3: An illustration of FissionVAE with Latent Space Decoupling. The latent variables are forced to follow their respective group prior distributions. The model is aggregated the same way as the baseline FedVAE.

When decoupling the latent space, the encoder maps the input data to different distributions based on the client's group. For instance, MNIST client may map to $\mathcal{N}(-1,1)$ and FashionMNIST clients to $\mathcal{N}(1,1)$. The KL divergence in the ELBO for this model is given by:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu_q},\boldsymbol{\sigma_q}||\mathcal{N}(\pm 1,1)) = \frac{1}{2}\Sigma_{i=1}^{k}[\sigma_i + \mu_i^2 \mp 2\mu_i - \log\sigma_i] \tag{1}$$

Here, $\boldsymbol{\mu_q}$ and $\boldsymbol{\sigma_q}$ represent the encoder's estimates for the parameters of the latent code's distribution, and k is the dimension of the latent code.

Figure 1 shows randomly generated amples produced after training the FissionVAE with latent space decoupling on the Mixed MNIST dataset. While the quality of reconstructed images are improved compared to the baseline FedVAE, the generated images still exhibit a mixture of handwritten digits and clothing items, even when explicitly sampling from their respective latent distributions. This suggests that while decomposing latent encoding helps improving reconstructions,

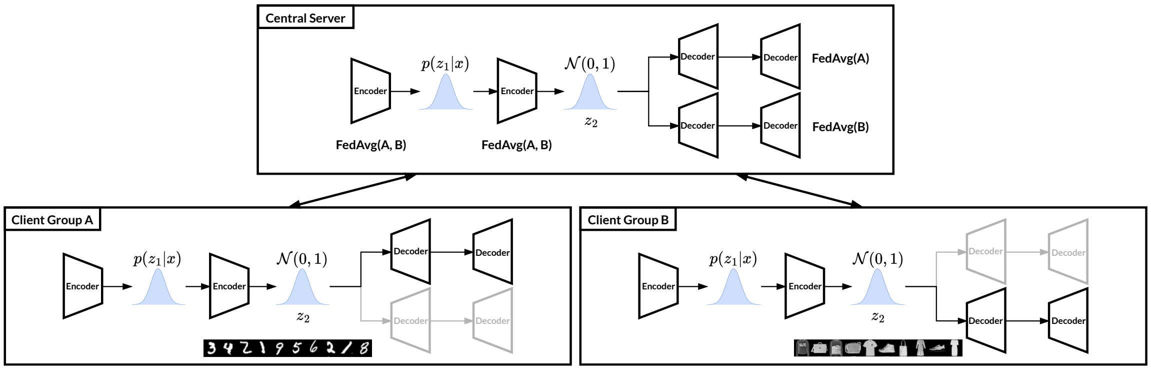

Figure 4: An illustration of Hierarchical FissionVAE. This FissionVAE architecture extends to allow two levels of latent variables. The latent variable $\boldsymbol{z_1}$ can be either learned or predefined. As input from different groups has been separated by $\boldsymbol{z_1}$, the latent variable $\boldsymbol{z_2}$ is set to follow the standard normal distribution.

the unified decoder still blends features due to the aggregation of model weights from diverse visual domains. This observation motivates the architecture described in the next section, where the decoder is also split based on client groups.

2.3 FissionVAE with Group-specific Decoder Branches

Non-Hierarchical FissionVAE Building on the concept introduced by FissionVAE with latent space decoupling, we further refines non-IID data generation by incorporating decoder branches specific to each data group while maintaining a unified encoder. This design allows the central server to aggregate the encoder updates agnostically of the client groups, whereas decoder branches are aggregated specifically according to their corresponding groups. In addition, this approach also offers flexibility in the choice of the prior latent distribution $p(\boldsymbol{z})$ for each group to exert more explicit control over the data generation through the decoder. Figure 5 illustrates this branching architecture.

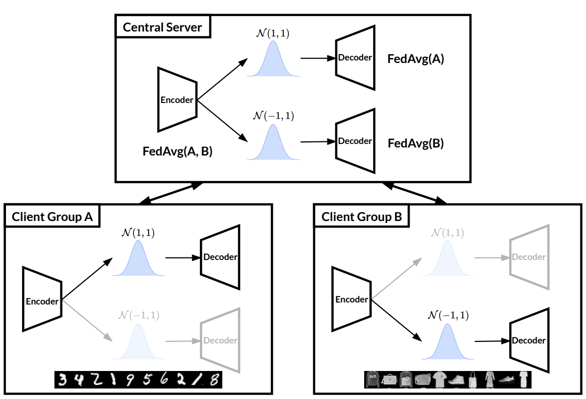

Figure 5: An illustration of FissionVAE with Decoder Branch Decoupling. This FissionVAE creates decoders specific to client groups and enforces constraints for latent variable priors. The encoder is aggregated across groups while the group-specific decoder is only aggregated from local models within the corresponding group.

Figure 1 also includes randomly generated samples produced after training the FissionVAE with decoder branches. The results indicate a significant reduction in the blending feature issue in previously discussed VAE architectures.

Hierachical FissionVAE Next, we show that the branching architecture can be enhanced by integrating hierarchical inference [Kingma *et al.*, 2016] [Sønderby *et al.*, 2016] to the federated learning framework, which enables the use of deeper network structures to capture more complex data distributions. Fig 4 depicts the FissionVAE with two levels of hierarchical inference. In this architecture, the first encoder module estimates $q(\boldsymbol{z_1}|\boldsymbol{x})$ from the input data, then the second encoder module estimates $q(\boldsymbol{z_2}|\boldsymbol{z_1})$ based on the first level latent code. The decoder reverses the encoding process, which estimates $p(\boldsymbol{z_1}|\boldsymbol{z_2})$ based on $\boldsymbol{z_2}$ to reconstruct $\boldsymbol{z_1}$, and subsequently reconstructs the original input $\boldsymbol{x}$ by estimating $p(\boldsymbol{x}|\boldsymbol{z_1})$.

Following the convention in hierarchical VAEs, we assume conditional independence among the latent codes.Then the ELBO for this hierarchical VAE is expressed as (refer to supplementary material for derivation),

$$\begin{aligned}\mathrm{ELBO}_H = &\ \mathbb{E}_{q_\phi(\boldsymbol{z_1}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z_1})] \\ &- \mathbb{E}_{q_\phi(\boldsymbol{z_1}|\boldsymbol{x})}[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1})||p(\boldsymbol{z_2}))] \\ &- \mathbb{E}_{q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1})}[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z_1}|\boldsymbol{x})||p_\theta(\boldsymbol{z_1}|\boldsymbol{z_2})] \qquad (2)\end{aligned}$$

In the equation above, the first term is the reconstruction term as it is the expectation of the log-likelihood for the input samples under the distribution estimated from the encoded $\boldsymbol{z_1}$, the second term is the prior matching term which is enforcing the encoded $\boldsymbol{z_2}$ to conform the prior distribution $\boldsymbol{z_2} \sim \mathcal{N}(0, 1)$, and the last term is the consistency term which requires $\boldsymbol{z_1}$ from either the encoder or the decoder to be consistence. In practice, we find that adding the reconstruction loss from $\boldsymbol{z_2}$ to $\boldsymbol{x}$ is also crucial for generating meaningful samples. Optionally, perceptual losses such as the VGG loss [Ledig *et al.*, 2017] or the structural similarity index measure (SSIM) [Wang *et al.*, 2004] loss can be used to promote the fidelity of reconstructed images. However, no significant improvement is observed in our experiments. Therefore no perceptual loss is included in our implementation. The final loss function for the hierarchical and branching FissionVAE then becomes,

$$\mathcal{L} = \mathbb{E}_{q_\phi(\boldsymbol{z_1}|\boldsymbol{x})}[D_{\mathrm{KL}}(q_\phi(\boldsymbol{z_1}|\boldsymbol{z_x})||p(\boldsymbol{z_1}))] - \mathbb{E}_{q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1})}[\log p_\theta(x|\boldsymbol{z_1}, \boldsymbol{z_2})] - \mathrm{ELBO}_H \quad (3)$$

Here we minimize the KL divergence for $\boldsymbol{z_1}$ only when the prior distribution for $\boldsymbol{z_1}$ is explicitly defined, otherwise the model learns the latent distribution by itself.

The proposed hierarchical FissionVAE also allows heterogeneous decoder architectures for each client groups, as each decoder branch is trained and aggregated independently. This flexibility is particularly advantageous in federated learning environments, where clients often possess varying computational resources. Client groups with more resources can implement deeper and more complex network structures, while groups with limited computational capacity can utilize lighter models.

Complexity of FissionVAE FissionVAE's space complexity grows linearly with the number of clients, due to group-specific decoder branches. Time complexity per client follows standard feedforward model training. While we use smaller batch sizes to encourage better latent space exploration, this does not change asymptotic complexity.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluated the proposed federated VAEs using two composite datasets. Mixed MNIST combines MNIST [LeCun and Cortes, 2010] and FashionMNIST [Xiao *et al.*, 2017], dividing samples into two client groups (one per dataset) with 10 clients each. Training samples were evenly distributed within each group, and the default test sets served as evaluation benchmarks. An equal number of images were generated using the global model for comparison.

CHARM is a more diverse dataset combining five domains: Cartoon faces [Churchill, 2019], Human faces [Karras *et al.*, 2018], Animals [Xian *et al.*, 2019], Remote sensing images [Helber *et al.*, 2019], and Marine vessels [Gundogdu *et al.*, 2016], using preprocessed square images from Meta-Album for AwA2 and MARVEL. Images were resized to 32×32, and each domain was represented by 20 clients, with 20,000 images for training and 5,000 for evaluation. As with Mixed MNIST, the global model generated evaluation samples.

For Mixed MNIST, encoders and decoders used Multi-Layer Perceptrons (MLPs). On CHARM, encoders $q(\boldsymbol{z_1}|\boldsymbol{x})$ and decoders $p(\boldsymbol{x}|\boldsymbol{z_1})$ were convolutional, while $q(\boldsymbol{z_2}|\boldsymbol{z_1})$ and $p(\boldsymbol{z_1}|\boldsymbol{z_2})$ used MLPs. Client participation followed a Bernoulli distribution: $B(0.5)$ for Mixed MNIST and $B(0.25)$ for CHARM. Hyperparameters included learning rates of 1×10^{-3} (Mixed MNIST) and 1×10^{-4} (CHARM), with 70 and 500 training rounds, respectively. Clients performed 5 local epochs per round with a batch size of 32. Centralized settings used 70 epochs for Mixed MNIST and 250 for CHARM.

Evaluation metrics included Fréchet Inception Distance [Heusel *et al.*, 2017] and Inception Score [Salimans *et al.*, 2016] for generation quality, and the negative log-likelihood (NLL) of the ELBO for reconstruction performance. IS was computed using an ImageNet-pretrained Inception model [Szegedy *et al.*, 2016].

3.2 Results and Analysis

Here we present the following experiments: we first evaluate the overall generative performance of the proposed VAE architectures in both federated and centralized settings, then we explore strategies for encoding the prior distribution $p(\boldsymbol{z_1})$, and lastly we showcase the use of heterogeneous decoder architectures in our FissionVAEs. For experiments investigating different generation pathways of hierarchical VAEs and the effect of reconstruction losses, please refer to our supplementary material.

Overall Performance

The overall performance of the proposed FissionVAE models is summarized in Table 1, and generated examples are shown in Fig. 6. In addition to the FedVAE baseline, a Deep Convolutional GAN (DCGAN) [Radford *et al.*, 2016] trained via FedGAN [Rasouli *et al.*, 2020] is used for comparison. Since GAN does not directly model the likelihood of data, NLL is not evaluated for FedGAN. Also, FedGAN on CHARM suffers from severe mode collapse, therefore performance evaluation is not available on this dataset. Notably, the performance of all models on the CHARM dataset is less robust compared to the Mixed MNIST dataset. This discrepancy arises because the CHARM dataset, encompassing RGB images from diverse domains, presents a more complex and realistic federated learning scenario. The dataset's diversity, coupled with a lower local data availability and participation rate among clients, poses greater challenges to federated generative models.

Latent Space Decoupling vs Decoder Branches As shown in Table 1, both latent space decoupling and group-specific decoder branches improve image quality (lower FID, higher IS). Decoder branches alone yield larger gains, highlighting the negative impact of mixing decoders trained on non-IID data.

FissionVAE+L moderately improves upon FedVAE by partitioning the latent space by client group, helping the decoder better distinguish domain-specific features and reducing representation overlap. Fig. 6 shows that while FissionVAE+L enables group-specific sampling, shared decoder aggregation still causes artifacts such as blended features.

FissionVAE+D, with a unified encoder and domain-specific decoder branches, greatly reduces visual blending. The encoder functions like a routing module akin to Mixture-of-Experts, which directs inputs to group-specific latent distributions. As decoders remain distinct during aggregation, texture mixing is avoided, producing cleaner outputs (Fig. 6).

FissionVAE+L+D combines both latent space decoupling and decoder branches. As shown in Table 1, FissionVAE+L+D yields marginal gains on Mixed MNIST but outperforms FissionVAE+D on CHARM. Enforcing latent space decoupling yields different outcomes depending on the number of client groups. For Mixed MNIST (2 groups), the FID is lowered due to the extra latent constraints. However, as the

Model	Mixed MNIST						CHARM					
	Federated			Centralized			Federated			Centralized		
	FID ↓	IS ↑	NLL ↓	FID ↓	IS ↑	NLL ↓	FID ↓	IS ↑	NLL ↓	FID ↓	IS ↑	NLL ↓
FedGAN	118.52	2.39	-	91.08	3.18	-	-	-	-	-	-	-
FedVAE	117.03	2.29	0.23	40.59	**3.62**	**0.18**	167.18	1.57	40.80	89.26	2.57	46.99
FissionVAE+L	64.99	2.83	**0.22**	39.27	3.03	0.18	155.81	1.73	43.49	86.19	2.53	51.45
FissionVAE+D	**40.78**	3.01	0.26	34.76	3.05	0.25	120.39	2.16	33.07	63.25	**2.95**	36.76
FissionVAE+L+D	42.11	**3.04**	0.25	34.39	3.08	0.20	109.10	2.27	33.29	**50.30**	2.89	40.14
FissionVAE+H+L+D	47.72	2.98	0.30	**28.82**	3.16	0.24	**107.69**	**2.32**	**27.46**	74.59	2.58	**27.09**

Table 1: Evaluation of proposed FissionVAEs on the Mixed MNIST and CHARM dataset. +L is for decoupled latent space. +D is for branching decoders. +H is for the hierarchical architecture. Best results in are in **bold**. Second best results are underlined. ↑ denotes the higher the better, while ↓ means the lower the better.

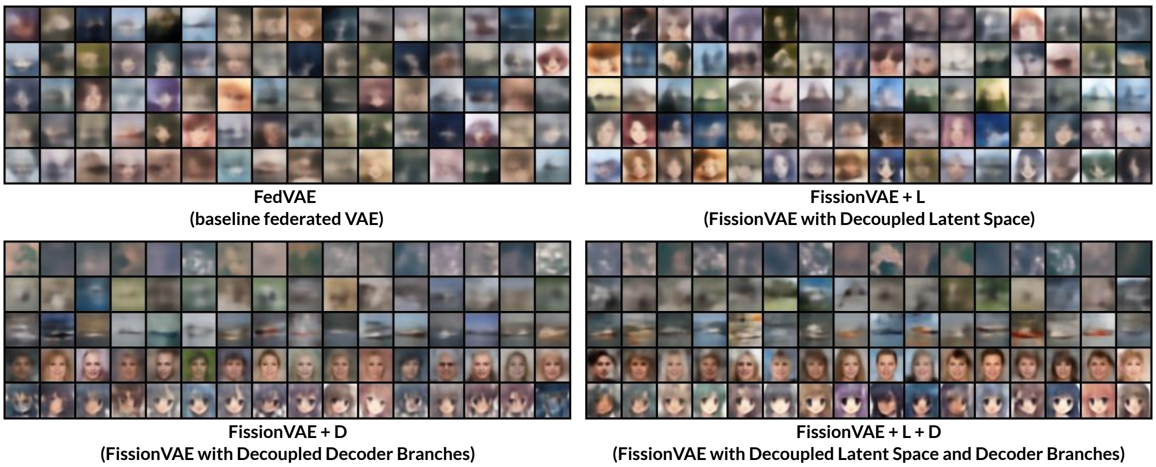

Figure 6: Qualitative results of image generation with FissionVAEs on the CHARM dataset. Best viewed in color.

number of client groups increases on CHARM (5 groups), explicit latent space decoupling provides more direct signal to the VAE to identify the intra-group difference, resulting an improved FID. In Fig. 6 it can be observed that images generated by FissionVAE+L+D are sharper than the ones generated by FissionVAE+D.

Hierarchical FissionVAE As discussed in Section 2, here we consider a hierarchical VAE with two levels of latent variable. In Table 1, the architecture FissionVAE+H+L+D performs the best on the CHARM dataset and falls behind its non-hierarchical counterpart on the Mixed MNIST dataset. The hierarchical VAE employs multiple levels of latent representations, which refines the model's ability to capture and reconstruct complex data distributions more faithfully. The performance degradation on simpler datasets like Mixed MNIST suggests that the hierarchical approach might introduce unnecessary redundancy without proportional gains in performance.

Decoupling the Prior of z_1

Explicitly decoupling the latent space for different client groups improves the ability of VAEs to generate images that align with the true data distribution (Table 1). We explore several priors for the latent distribution, modeled as multivariate Gaussians with customizable means and identity covariance matrices and evaluate them in Table 2. Details regarding the formal definition of priors can be found in the supplementary material.

Model	Prior $p(z_1)$	Mixed MNIST		CHARM	
		FID ↓	IS ↑	FID ↓	IS ↑
FissionVAE+L+D	identical	**40.78**	3.01	120.39	2.16
	one-hot	42.01	3.02	113.82	2.25
	symmetrical	41.79	2.95	-	-
	random	43.26	3.00	111.77	**2.47**
	wave	42.11	**3.04**	**109.10**	2.27
FissionVAE+H+L+D	identical	55.91	2.96	122.16	**2.30**
	one-hot	**53.22**	2.97	121.33	2.29
	symmetrical	58.21	**3.03**	-	-
	random	53.99	2.94	124.91	2.23
	wave	53.68	2.94	**118.56**	2.24
	learnable	47.72	2.98	107.69	2.32

Table 2: Evaluation of Generation Performance with z_1 Priors

In non-hierarchical VAEs, $\boldsymbol{z_1}$ represents the sole latent variable, while in hierarchical VAEs, $\boldsymbol{z_1}$ is controlled, with $\boldsymbol{z_2}$ following a standard normal distribution $N(0, 1)$. Baseline priors are identical across client groups. Other prior variations include one-hot encoding, symmetrical positive and negative integers, random vectors, wave encodings (with grouped 1's in dimensions corresponding to client groups), and a learnable approach unique to hierarchical VAEs. The learnable approach dynamically aligns priors but sacrifices

Decoder Architecture on the FashionMNIST Branch	MNIST			FashionMNIST			Overall		
	FID ↓	IS ↑	NLL ↓	FID ↓	IS ↑	NLL ↓	FID ↓	IS ↑	NLL ↓
Homogeneous	**46.73**	**2.41**	0.38	61.81	2.92	0.61	**47.72**	2.98	**0.30**
Deeper MLP	49.54	2.38	**0.33**	**60.95**	2.90	0.78	48.79	2.95	0.39
Deeper MLP + Conv	48.21	2.38	0.38	65.82	**2.99**	**0.60**	50.16	**3.00**	0.30

Table 3: Evaluation of FissionVAE+H+L+D with Heterogeneous Decoder Architectures on the Mixed MNIST

direct sampling from $p(\boldsymbol{z_1})$.

Hierarchical FissionVAE often underperforms non-hierarchical variants when predefined priors are used due to increased uncertainty from additional latent layers. However, the learnable approach excels in capturing complex distributions dynamically. In simpler datasets like Mixed MNIST, identical priors suffice, but explicit latent encoding becomes crucial as client group diversity increases, as seen with CHARM.

Among prior definitions, symmetrical priors often lead to divergence on CHARM, as their means may exceed neural network initialization ranges. One-hot and random approaches show comparable results but are less consistent than wave encoding, which clearly distinguishes group priors without out-of-range values.

Group-level Privacy

In the presence of hierarchical VAEs, it is possible to incorporate the encoder $q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1})$ into the generation process, that is, we can first sample the latent code $\boldsymbol{z_1}$ from its prior distribution, then feed it to the subsequent encoder $q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1})$ and the decoders $p_\theta(\boldsymbol{z_1}|\boldsymbol{z_2})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z_1})$ to obtain the synthesize a generated sample. On the Mixed MNIST dataset, we observe that swapping the prior distributions of the two client groups in the such a generation pathway leads to evident mode collapse, shown in Figure 7. This suggests that the group-level privacy may be preserved by maintaining the confidentiality of prior distributions. This strategy ensures that high-quality samples are generated only when the correct prior distribution is used, while mismatched distributions yield unrecognizable outputs. This phenomenon is more pronounced in both hierarchical and non-hierarchical FissionVAEs on the Mixed MNIST dataset than on the CHARM dataset, likely due to the simpler, more uniform nature of the Mixed MNIST data compared to the diverse and colorful image types in CHARM, which pose greater challenges in satisfying complex latent distribution constraints. Evaluation on other generation pathways are presented in the supplementary material.

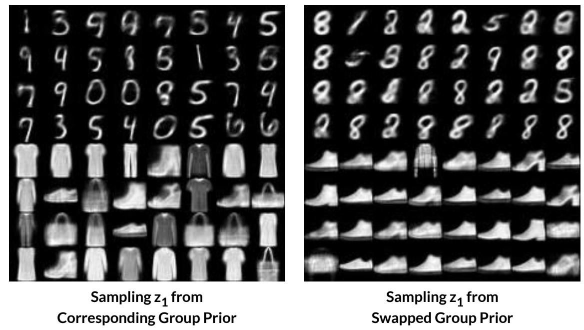

Figure 7: In hierarchical FissionVAE, when the prior distribution $p(\boldsymbol{z_1})$ of the MNIST and FashionMNIST groups are swapped, the generation pathway $q(\boldsymbol{z_1}) \rightarrow q_\phi(\boldsymbol{z_2}|\boldsymbol{z_1}) \rightarrow p_\theta(\boldsymbol{z_1}|\boldsymbol{z_2}) \rightarrow p_\theta(\boldsymbol{x}|\boldsymbol{z_1})$ leads to sever mode collapse, suggesting potential group-level privacy preserving through protected prior distribution.

Heterogeneous Decoders in FissionVAE

As discussed in Section 2, the decoupling of decoders for client groups allow for the use of heterogeneous architectures in FissionVAE. The Mixed MNIST dataset, with its relatively simple and grayscale colors, can be generated from both fully connected (MLP) and convolutional layers. In contrast, the more complex and colorful images in the CHARM dataset predominantly require convolutional layers for effective generation.

Table 3 details the performance evaluation of various decoder architectures. The term 'homogeneous' refers to identical architectural configurations across all decoder branches, namely a three-layer MLP for each decoder modules. In the 'Deeper MLP' configuration, we add two additional fully connected layers to both $p_\theta(\boldsymbol{z_1}|\boldsymbol{z_2})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z_1})$. Meanwhile, we completely replace the decoder $p_\theta(\boldsymbol{x}|z_1)$ from MLP to a series of transpose convolution layers in the 'Deeper MLP + Conv' configuration. The results indicate a gradual reduction in overall FID scores as the decoder architecture becomes more heterogeneous. However, the integration of convolutional layers does not improve generation performance over the MLP models, underscoring that while heterogeneous architectures are feasible, they can disrupt the convergence of the VAE due to mismatches in architecture and the model's weight space.

4 Conclusion

We presented FissionVAE, a generative model for federated image generation in non-IID data settings. By decoupling the latent space and employing group-specific decoder branches, FissionVAE enhances generation quality while preserving the distinct features of diverse data subsets. Experiments on Mixed MNIST and CHARM datasets demonstrated significant improvements over baseline federated VAE models, with heterogeneous decoder branches and wave-encoded priors proving particularly effective. Future work includes improving the stability of heterogeneous decoder branches, enabling cross-modality data generation, and developing scalable strategies for handling an increasing number of client groups in real-world federated learning scenarios.

Acknowledgments

This work is supported by the EPSRC National Edge AI Hub (EP/Y007697/1).

References

[Bohacek and Farid, 2023] Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse, 2023.

[Borsos *et al.*, 2023] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. In *arXiv*, 2023.

[Chen and Vikalo, 2023] Huancheng Chen and Haris Vikalo. Federated learning in non-iid settings aided by differentially private synthetic data. In *CVPRW*, 2023.

[Churchill, 2019] Spencer Churchill. Anime face dataset, 2019.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NIPS*, 2014.

[Gundogdu *et al.*, 2016] Erhan Gundogdu, Berkan Solmaz, Veysel Yucesoy, and Aykut Koc. Marvel: A large-scale image dataset for maritime vessels. In *Asian Conference on Computer Vision*, 2016.

[Hardy *et al.*, 2019] Corentin Hardy, Erwan Le Merrer, and Bruno Sericola. Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2019.

[Heinbaugh *et al.*, 2023] Clare Elizabeth Heinbaugh, Emilio Luz-Ricca, and Huajie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *ICLR*, 2023.

[Helber *et al.*, 2019] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[Jiang *et al.*, 2023] Yuchen Jiang, Ying Wu, Shiyao Zhang, and James J.Q. Yu. Fedvae: Trajectory privacy preserving based on federated variational autoencoder. In *IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, 2023.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[Karras *et al.*, 2020] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NIPS*, 2020.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Kingma *et al.*, 2016] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016.

[LeCun and Cortes, 2010] Yann LeCun and Corinna Cortes. MNIST handwritten digit database, 2010.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[Li *et al.*, 2025] Zhiwei Li, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. Personalized federated collaborative filtering: A variational autoencoder approach. In *AAAI*, 2025.

[McMahan *et al.*, 2023] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023.

[Polato, 2021] Mirko Polato. Federated variational autoencoder for collaborative filtering. In *2021 International Joint Conference on Neural Networks*, 2021.

[Radford *et al.*, 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[Rasouli *et al.*, 2020] Mohammad Rasouli, Tao Sun, and Ram Rajagopal. Fedgan: Federated generative adversarial networks for distributed data. In *arXiv:2006.07228*, 2020.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[Shumailov *et al.*, 2024] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 2024.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[Sønderby *et al.*, 2016] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *NIPS*, 2016.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. In *arXiv*, 2023.

[Wang *et al.*, 2004] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[Xian *et al.*, 2019] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. In *arXiv:1708.07747*, 2017.

[Xiong *et al.*, 2023] Zuobin Xiong, Wei Li, and Zhipeng Cai. Federated generative model on multi-source heterogeneous data in iot. In *AAAI*, 2023.

[Yonetani *et al.*, 2019] Ryo Yonetani, Tomohiro Takahashi, Atsushi Hashimoto, and Yoshitaka Ushiku. Decentralized learning of generative adversarial networks from non-iid data. In *CVPR Workshop on Challenges and Opportunities for Privacy and Security*, 2019.

[Zhang *et al.*, 2024] Lu Zhang, Qian Rong, Xuanang Ding, Guohui Li, and Ling Yuan. Efvae: Efficient federated variational autoencoder for collaborative filtering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024.