

Guiding Diffusion Models with Fine-Grained Conditions and Semantics-Preserving Sampling for One-Shot Federated Learning

Xiaojun Deng¹ Tianchi Liao² Zhiyuan Liu¹ Chuan Chen^{1*} Zibin Zheng²

¹School of Computer Science and Engineering, Sun Yat-sen University

²School of Software Engineering, Sun Yat-sen University

dengxj26@mail2.sysu.edu.cn, chenchuan@mail.sysu.edu.cn

Abstract

*One-shot Federated Learning (OSFL) has emerged as a promising paradigm to mitigate the high communication overhead of traditional federated learning. However, its effectiveness is often hindered by data heterogeneity across clients. While recent methods leverage pre-trained diffusion models to generate data for OSFL, they often struggle with some practical limitations, including a lack of semantic fidelity in capturing the fine-grained characteristics of local data, and insufficient diversity in the generated data, which collectively degrade the performance of the global model. To address these challenges, we propose Espresso, a novel framework that enhances both the fidelity and diversity of synthetic data in OSFL. Espresso consists of two main components: (1) **Fine-Grained Condition Learning**, which learns fine-grained conditional embeddings to improve semantic fidelity and diversity by modeling intra-category patterns, and (2) **Semantics-Preserving Sampling**, which diversifies the generated data by modeling the distribution of latent noises and applying a self-reflection sampling strategy. Extensive experiments on benchmark datasets demonstrate that Espresso can improve the semantic fidelity and diversity of the synthetic data, leading to an enhancement in the performance of the global model compared to state-of-the-art OSFL methods under data heterogeneity.*

1. Introduction

Federated Learning (FL) [2, 25, 32] has emerged as a popular paradigm for collaborative machine learning, enabling multiple parties to train a shared model without exposing their private data [35]. The foundational framework for FL is established by iterative methods (e.g., Federated Averaging, FedAvg [32]). In each round, a central server distributes the current parameters of the global model to se-

lected clients. The clients then perform local training on their private data before uploading their updated models, which are then aggregated by the server to produce a new global model [32, 46]. However, this process is hindered by several challenges. Firstly, it is vulnerable to data heterogeneity [24, 26, 27, 36], or non-IID (non-Independent and Identically Distributed) data, which mainly manifests as either label skew, where clients hold different distributions of categories, or feature skew, where the same category exhibits different feature distributions across clients. In addition, the volume and quality of each client's local data vary across clients in some practical settings, which may also affect the performance [2]. Furthermore, numerous rounds of model exchange may cause a large communication overhead, which may be unacceptable in some practical settings.

To reduce communication costs, One-Shot Federated Learning (OSFL) [11, 31, 56] is proposed to limit communication to a single round. Some existing OSFL methods [1, 9, 56] involve clients to upload their local models, which are then aggregated through techniques such as ensembling or knowledge distillation [14]. However, knowledge from local models becomes increasingly difficult to consolidate as data heterogeneity becomes severe. [31] In addition to utilizing local models, another possible approach is to use generative approaches for global model training. Some methods [41, 60] propose to distill local data into a small-scale compressed dataset by data distillation [45], which is efficient for communication but still suffers from the information loss caused by compression. Some methods [6, 56] employ earlier generative models such as GANs [10] which often fall short in the fidelity and diversity of synthetic data.

Recently, to address the above challenges arising in OSFL, some methods [3, 48–50] leverage the generative power of pre-trained diffusion models [16, 38, 47] by guiding them to produce synthetic data that approximate the distribution of clients' local data. The global model is then trained on these synthetic data. Pre-trained diffusion models are trained on large-scale general datasets, thereby possessing the capacity to generate images of most styles and

*Corresponding Author.

objects with conditional guidance information [29]. Therefore, in FL settings, it provides the possibility of reproducing the distribution of client data with less loss of information on the server [50]. A key advantage of this approach is that the global model can now more accurately observe the commonalities and differences among different client distributions, thereby promoting the performance of generalization. However, these methods still suffer from some limitations: *The guidance information transmitted by clients is often too simplistic to capture the rich patterns of local data, especially the intra-category differences*, leading to low semantic fidelity and diversity where the synthetic data cannot sufficiently reflect the distribution of local data.

To overcome these limitations, we propose Espresso, a novel method which is based on pre-trained diffusion models and designed to enhance both the semantic fidelity and diversity of synthetic data in OSFL. Espresso tackles the issue of insufficient semantic fidelity and diversity through Fine-Grained Condition Learning, which learns fine-grained conditional embeddings by identifying and modeling intra-category patterns within each client’s dataset. Moreover, to further align the synthetic data with the conditional embeddings, we also propose Semantics-Preserving Sampling strategy. First, we model the distribution of the latent noises at the final diffusion step as Gaussian Mixture Models, from which we sample the denoising starting point to ensure that the denoising starting point is variable and meaningful. Then, we apply Z-Sampling [28], a self-reflection strategy, to iteratively optimize the latent vectors in a zig-zag manner during the denoising process, better injecting the information from conditional embeddings. In essence, Espresso utilizes compressed but representative information from clients to guide a diffusion model, thereby reconstructing the client’s data distribution on the server for better global model. This process is analogous to brewing a cup of coffee from concentrated espresso powder. The experiments demonstrate that Espresso can enhance the semantic fidelity and diversity of the synthetic data and further the performance of the global model, validating the effectiveness of Espresso in heterogeneous settings. The main contributions of this work are as follows:

- To improve the semantic fidelity and diversity of synthetic data in existing diffusion-based OSFL methods, we propose Fine-Grained Condition Learning to further capture the characteristics of intra-category patterns.
- To further enhance the semantic alignment of synthetic data with the conditional embeddings, we employ Gaussian Mixture Models to model the latent noises in the diffusion process as starting points of denoising, and Z-Sampling strategy to better inject conditional information to the latent vectors.
- We empirically validate our approach on several benchmark datasets, demonstrating significant performance im-

provements over state-of-the-art OSFL methods in most heterogeneous settings.

2. Related Works

2.1. One-shot Federated Learning

Some traditional FL methods [18, 26, 33] have been designed to handle different types of data heterogeneity. Although effective, this paradigm still incurs substantial communication overhead. One-shot Federated Learning (OSFL) addresses this by limiting communication to a single round. Some OSFL methods require clients to upload local models for server-side aggregation by ensembling [1, 42, 44], prototype learning [43, 53], knowledge distillation [4, 9, 23] and Bayesian methods [19, 30]. However, under higher data heterogeneity, the effectiveness of these methods decreases as divergent local models are more difficult to integrate. Therefore, an alternative paradigm has emerged to allow the clients to transmit a representation of their data distributions. Some methods [41, 57, 60] achieve this by dataset distillation [45, 52], GANs [6, 21, 54, 56] or VAEs [13, 57]. Recently, some methods [3, 49, 50] require clients to upload some form of guidance information that is used by the server to guide a diffusion model to generate a synthetic dataset for global model training. However, most of such guidance information is not enough to precisely capture the distribution of local data, especially the intra-category patterns, which limits the performance of global model. Espresso handles this by discovering different patterns within a category, by which we can learn an individual conditional embedding for each pattern.

2.2. Diffusion Models

The theoretical foundations of diffusion models are first established by DDPM [16] and DDIM [40]. Their practical application to high-resolution image synthesis was significantly advanced by Latent Diffusion Models (LDMs) [38], which perform the diffusion process in a compressed latent space. Pre-trained diffusion models such as Stable Diffusion [7, 38] further enabled controllable generation by leveraging powerful pre-trained multi-modal models, *e.g.*, CLIP [37], as a conditioning mechanism. Based upon this, some diffusion personalization techniques are proposed to align a pre-trained diffusion model to fit a specific dataset, such as Textual Inversion [8], DreamBooth [39], LoRA [17], ControlNet [58] and IP-Adapter [51], by learning some guidance information to represent the underlying visual concepts in the dataset. Textual Inversion is noted for its parameter efficiency, making it suitable for FL settings with its low training, communication, and storage overhead. However, due to the fact that Textual inversion learns token embeddings, which need to be combined with natural language templates and thus limit their ability to learn less common

styles. Therefore, in Espresso, we directly bypass the text encoder and learn the entire conditional embedding, which has more learning ability in dealing with such styles.

3. Preliminaries

3.1. Federated Learning

The primary objective of Federated Learning is to minimize a global loss function, which is a weighted average of the local loss functions of clients [2, 25]. Given the local dataset of client k as $\mathcal{D}_k$ and the global dataset as $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$, the global objective can be formulated as:

$$\min_{\omega} \mathcal{L}(\omega; \mathcal{D}) = \sum_{k=1}^K p_k \mathcal{L}_k(\omega; \mathcal{D}_k), \quad (1)$$

where K is the number of clients, p_k is the weight for client k , and $\mathcal{L}_k(\omega; \mathcal{D}_k)$ is the local loss function with model parameters ω . However, due to privacy limitations, the server can not directly compute $\mathcal{L}_k(\omega; \mathcal{D}_k)$. Therefore, traditional FL methods [26, 32] try to minimize $\mathcal{L}(\omega; \mathcal{D})$ in an iterative paradigm. In contrast, generative models, especially diffusion models, present an alternative approach for the server to approximate $\mathcal{L}_k(\omega; \mathcal{D}_k)$ with high-quality synthetic data.

3.2. Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) [16] are typical diffusion models that generate data by reversing a noising process. Typically, we consider a conditioned latent diffusion model guided by additional information, such as class labels or text embeddings.

Forward Diffusion Process. The forward process gradually adds Gaussian noise to the latent $z^0 = \mathcal{E}(x^0)$ of data point x^0 over T timesteps:

$$q(z^t | z^{t-1}) = \mathcal{N}(z^t; \sqrt{1 - \beta_t} z^{t-1}, \beta_t I), \quad (2)$$

where β_t is a small fixed positive constant representing the noise schedule at timestep t . A property of this process is that z^t can be directly sampled from z^0 :

$$z^t = \sqrt{\alpha_t} z^0 + \sqrt{1 - \alpha_t} \epsilon, \quad (3)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, and $\epsilon \sim \mathcal{N}(0, I)$.

Reverse Denoising Process. The reverse process learns to restore the original latent from pure noise z^T . Given condition c , a neural network ϵ_θ is trained to predict the noise added at each step. Given the predicted noise $\epsilon_\theta(z^t, t, c)$, the latent for the previous timestep z^{t-1} is:

$$z^{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z^t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(z^t, t, c) \right) + \sigma_t \epsilon_t, \quad (4)$$

where ϵ_t is Gaussian noise and σ_t is a variance term.

Training Objective. Using reparameterization trick, the simplified objective is derived as:

$$\min_{\theta} \mathbb{E}_{t, z^0, c, \epsilon} [\| \epsilon - \epsilon_\theta(\sqrt{\alpha_t} z^0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, c) \|^2]. \quad (5)$$

It is worth noting that given a pre-trained θ and target original data, condition c can also be optimized with similar objective for generating data similar to the original data.

3.3. DDIM and DDIMInverse

Denoising Diffusion Implicit Models (DDIMs) [40] learn a deterministic and non-Markovian process for faster denoising than DDPM. The denoising step at t is:

$$z^{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z^t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z^t, t, c)}{\sqrt{\bar{\alpha}_t}} + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon_\theta(z^t, t, c) \right). \quad (6)$$

DDIMInverse is the inverse process of DDIM. Given a data point z^0 , DDIMInverse approximates the forward diffusion process to find a corresponding deterministic latent vector z^t , which can be easily reconstructed to z^0 by DDIM.

3.4. Classifier-Free Guidance

Classifier-Free Guidance (CFG) [15] improves conditional sample quality without an external classifier, a technique adopted in most current diffusion models. ϵ_θ is trained with both condition c and null condition $\emptyset$. During inference, the actual estimated noise at t is defined as:

$$\hat{\epsilon}_\theta(z^t, t, c) = \epsilon_\theta(z^t, t, \emptyset) + \gamma \cdot (\epsilon_\theta(z^t, t, c) - \epsilon_\theta(z^t, t, \emptyset)), \quad (7)$$

where $\gamma > 1$ is the guidance scale representing the conditioning effect.

4. Proposed Approach

The overview of Espresso is presented in Fig. 1. It consists of three main phases: (1) locally executed Fine-Grained Condition Learning, (2) global data generation with Semantics-Preserving Sampling, which is supported by GMM of noisy latent vectors and Z-sampling, and (3) Global Model Training with synthetic data. We begin by illustrating Fine-Grained Condition Learning.

4.1. Fine-Grained Condition Learning

The existing diffusion-based OSFL methods [3, 50] only learn a single piece of conditional information for each category. However, data belonging to the same category may exhibit various visual patterns that cannot be fully captured by it. To address this, the first stage of Espresso is to identify fine-grained intra-category patterns and learn a descriptive conditional embedding for each of them, as shown in ① of Fig. 1.

To discover the intra-category patterns, we first extract the representations of data for further analysis. For each client k , define its local dataset as $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{N_k}$, where x_i is an image and y_i is its corresponding label. We employ a powerful pre-trained vision encoder Φ (e.g.,

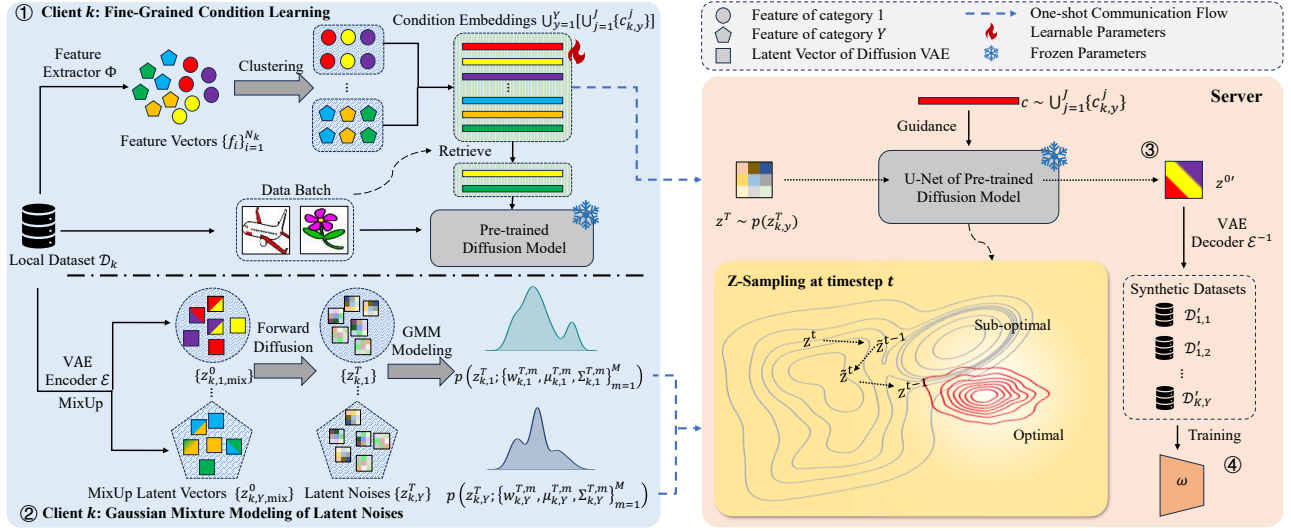


Figure 1. An overview of the Espresso. ① Fine-Grained Condition Learning identifies intra-category clusters of local dataset to learn fine-grained conditional embeddings, which enhances the semantic fidelity and diversity of synthetic dataset. ② The first part of Semantics-Preserving Sampling: GMM of Noisy Latent Vectors models the distribution of latent vectors z^T for better denoising starting point. ③ The second part of Semantics-Preserving Sampling: Guided with conditional embedding, the server samples an initial noise from the GMM and iteratively refining the generation path using Z-Sampling during denoising. ④ The final global model is trained on the synthetic dataset.

a CLIP Image Encoder [37] or DINOv2 [20]) to extract a representation vector $f_i = \Phi(x_i)$ for every image $x_i \in \mathcal{D}_k$. Then, we excavate the intra-category pattern by clustering. For each category y , we apply clustering to the set of extracted feature vectors $\{f_i | y_i = y\}$. This partitions the data of category y into several distinct clusters, $\{\mathcal{D}_{k,y}^j\}_{j=1}^J$, where J is the number of clusters. Each cluster $\mathcal{D}_{k,y}^j$ now represents a visually similar subgroup within category y .

After identifying the semantic clusters, we leverage a pre-trained conditional diffusion model to condense the clusters into light-weight conditional embeddings. Instead of using the pre-trained text encoder, we directly optimize the conditional embedding $c_{k,y}^j$, which is fed into the cross-attention layers of the diffusion model’s U-Net. This provides a better expressive capacity to capture the visual characteristics of each cluster. In this process, the parameters θ of the pre-trained diffusion model ϵ_θ are frozen, and only the conditional embedding $c_{k,y}^j$ is optimized. The objective for $c_{k,y}^j$ is the same as Eq. (5):

$$\min_{c_{k,y}^j} \mathbb{E}_{t, x \sim \mathcal{D}_{k,y}^j, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t} \mathcal{E}(x) + \sqrt{1 - \alpha_t} \epsilon, t, c_{k,y}^j)\|^2 \right]. \quad (8)$$

By optimizing Eq. (8), $c_{k,y}^j$ is trained to guide the diffusion model to denoise towards data that match the specific style and semantics of $\mathcal{D}_{k,y}^j$. Although it involves more learnable parameters per category, the computational and communication cost remains manageable and yields more descriptive guidance information. At the end of this stage, each client

k uploads all conditional embeddings $\bigcup_{y=1}^Y \left[\bigcup_{j=1}^J \{c_{k,y}^j\} \right]$ to the server for guiding the same diffusion model.

4.2. Semantics-Preserving Sampling

In standard diffusion denoising, the initial latent noise is drawn from standard Gaussian distribution, i.e., $z^T \sim \mathcal{N}(0, I)$. However, recent works suggest that some initial noise can generate data that leans towards certain visual concepts [12], and the reflection strategy during denoising can also improve the quality of synthetic data by accumulating more conditional semantic information [28]. To leverage these insights in OSFL, we propose to model latent noise using Gaussian Mixture Models and refine the process of denoising with self-reflection strategy, which collectively support the Semantics-Preserving Sampling of Espresso.

4.2.1. Gaussian Mixture Modeling of Latent Noises

For category y of client k , a natural approach is to model the distribution of z^T , which can be used to sample new initial latent noises. We can first model the distribution of z^0 as a Gaussian Mixture Model (GMM) with M components:

$$p(z_{k,y}^0) = \sum_{m=1}^M w_{k,y}^{0,m} \mathcal{N}(z_{k,y}^0 | \mu_{k,y}^{0,m}, \Sigma_{k,y}^{0,m}), \quad (9)$$

where $w_{k,y}^{0,m}$, $\mu_{k,y}^{0,m}$, and $\Sigma_{k,y}^{0,m}$ are the weight, mean, and covariance of the m -th component, respectively. Given that $z^T = \sqrt{\alpha_T} z^0 + \sqrt{1 - \alpha_T} \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, the distri-

bution of z^T can also be derived as a GMM:

$$p(z_{k,y}^T) = \sum_{m=1}^M w_{k,y}^{0,m} \mathcal{N}(z_{k,y}^T | \sqrt{\bar{\alpha}_T} \mu_{k,y}^{0,m}, \bar{\alpha}_T \Sigma_{k,y}^{0,m} + (1 - \bar{\alpha}_T) I). \quad (10)$$

However, directly modeling $p(z_{k,y}^0)$ presents practical challenges. The dimension of $z_{k,y}^0$ is too high and computationally hard to model using full covariance matrices. A simplification is to use diagonal covariance matrices, but it fails to capture correlations between different coordinates of latent space, ultimately limiting the diversity and quality of sampled $z_{k,y}^T$. To address this, we propose an alternative strategy. Instead of modeling the complex $p(z_{k,y}^0)$ and then $p(z_{k,y}^T)$, we directly model $p(z_{k,y}^T)$, which is, by design, much closer to $\mathcal{N}(0, I)$. This makes its spatial distribution more concentrated and thus more amenable to be modeled as a GMM with diagonal covariance matrices.

As shown in ② of Fig. 1, to approximate the process of sampling from the hypothetical (but intractable) $p(z_{k,y}^0)$, we first generate a batch of augmented latent vectors $z_{k,y,\text{mix}}^0$ via MixUp augmentation [55] on latent vectors from $z_{k,y}^0$. Then, each of $z_{k,y,\text{mix}}^0$ is diffused to $z_{k,y,\text{mix}}^T$ multiple times, each time with a different $\epsilon \sim \mathcal{N}(0, I)$. We collect all $z_{k,y,\text{mix}}^T$ and fit a GMM to them as an approximation to derived $p(z_{k,y}^T)$, whose parameters $w_{k,y}^{T,m}$, $\mu_{k,y}^{T,m}$ and $\Sigma_{k,y}^{T,m}$ are uploaded to the server along with learned conditional embeddings in Sec. 4.1. When the server needs to generate data of client k and category y , it first randomly selects a conditional embedding $c_{k,y}^j$ and then samples an initial noise z^T from the corresponding GMM $p(z_{k,y}^T)$.

4.2.2. Z-Sampling for Semantic Enhancement

Once the server has sampled $z^T \sim p(z_{k,y}^T)$ and selected a $c_{k,y}^j$, it denoises z^T to a data point guided by $c_{k,y}^j$. In this process, classifier-free guidance (CFG) is used to steer the generation process towards the desired semantics. The predicted noise at each timestep t is the same as Eq. (7):

$$\hat{\epsilon}_\theta(z^t, t, c_{k,y}^j) = \epsilon_\theta(z^t, t, \emptyset) + \gamma \cdot (\epsilon_\theta(z^t, t, c_{k,y}^j) - \epsilon_\theta(z^t, t, \emptyset)). \quad (11)$$

To further inject the information of $c_{k,y}^j$ to the synthetic data, we further employ Z-Sampling [28], as shown in ③ of Fig. 1. Z-Sampling is a self-reflection technique that iteratively refines the latent vectors during denoising to inject more semantic information. Instead of a monotonic path from z^T to z^0 , Z-Sampling introduces a “zigzag” operation. When t is larger than a threshold T_z , z^t is first denoised to obtain $\tilde{z}^{t-1}$. This step uses a stronger guidance scale γ_1 to push the latent towards the target condition $c_{k,y}^j$. Immediately following this, a deterministic DDIMinverse step is performed to map $\tilde{z}^{t-1}$ back to a new latent $\tilde{z}^t$ with a weaker guidance scale $\gamma_2 < \gamma_1$. Then, the true denoising step is performed by denoising $\tilde{z}^t$ to z^{t-1} again with γ_1 .

This zigzag process leverages the gap $\delta_\gamma = \gamma_1 - \gamma_2$, which means that some semantic information can be injected into $\tilde{z}^t$ before the true denoising step. By repeatedly applying it, the information contained in the conditional embeddings can be more sufficiently utilized, enhancing the semantic fidelity of the generated data. Finally, the denoised latent vector $z^{0'}$ is mapped back to image space with the VAE decoder $\mathcal{E}^{-1}$ of the pre-trained diffusion model.

4.3. Global Model Training

The final stage is the training of a global model with the synthetic dataset. Defining the synthetic dataset of client k is $\mathcal{D}'_k$, which consists of synthetic data of all categories. With the collection $\{\mathcal{D}'_k\}_{k=1}^K$, the objective of optimizing global model parameterized by ω is analogous to Eq. (1):

$$\min_{\omega} \mathcal{L}(\omega) = \sum_{k=1}^K p_k \mathbb{E}_{\mathcal{D}'_k} [\mathcal{L}_k(\mathcal{D}'_k)]. \quad (12)$$

By training with the synthetic dataset, the global model can learn the distributions of clients, without ever accessing their local datasets, and perform well on the complex and heterogeneous data distributions of clients, effectively overcoming the challenges of data heterogeneity.

5. Experiments

5.1. Setups

Datasets. We conduct experiments on both feature-skew and label-skew scenarios. For the feature-skew setting, we utilize following datasets: (1) **DomainNet** [34] comprises six domains: *clipart* (C), *infograph* (I), *painting* (P), *quickdraw* (Q), *real* (R) and *sketch* (S). We select 10 categories following [3]. (2) **PACS** [22] comprises four domains: *art* (A), *cartoon* (C), *photo* (P) and *sketch* (S). (3) **Common NICO++** [59] comprises six domains: *autumn* (A), *dim* (D), *grass* (G), *outdoor* (O), *rock* (R) and *water* (W). We randomly select 10 categories. We select 10 classes from Common NICO++, and for all three datasets, we assign a single domain to each client. To simulate realistic scenarios where clients may possess limited data, all experiments in these settings are conducted in a 16-shot per category. Further details about above datasets and that of datasets under the label-skew setting are provided in Appendix.

Baselines. We compare Espresso with the following baselines: (1) **Traditional FL methods** including FedAvg [32], FedProx [26], FPL [18] and FedSR [33], among which FPL and FedSR are designed for feature-skew settings by improving the representation learning. (2) **Non-Diffusion OSFL methods**, including Fens [1] where clients upload their local models for ensembling; FedD3 [41] where clients condense local datasets by dataset distillation; and DENSE [56] where server utilizes the local models to

Table 1. Performance comparison on DomainNet and Common NICO++ datasets under the 16-shot setting. All values are accuracy (%). The best and second performance in each column (excluding the theoretical upper-bound *Ceiling*) are in **bold** and underline.

Methods	DomainNet							Common NICO++						
	C	I	P	Q	R	S	Avg.	A	D	G	O	R	W	Avg.
<i>Ceiling</i>	85.25	58.69	78.93	83.76	90.48	83.84	80.16	92.69	84.41	91.82	88.75	84.91	78.27	86.81
FedAvg	76.44	46.63	63.09	56.36	87.14	71.76	66.90	79.24	63.61	83.76	76.04	76.49	68.54	74.61
FedProx	77.78	46.83	64.37	55.82	86.85	71.12	67.13	78.95	65.59	83.15	75.14	75.79	70.36	74.83
FPL	80.74	48.17	67.64	56.15	88.01	75.08	69.30	79.82	70.30	81.24	74.77	75.79	71.73	75.61
FedSR	<u>80.93</u>	49.69	68.69	54.06	87.92	74.30	69.27	78.07	67.82	84.00	75.86	78.95	71.28	75.99
DENSE	76.03	48.79	65.43	51.94	84.30	68.19	65.78	74.13	57.27	78.34	68.70	72.81	62.27	68.92
FedD3	66.55	51.49	63.30	56.30	78.88	63.64	63.36	70.13	61.91	70.80	62.17	65.29	60.93	65.21
Fens	71.65	47.85	63.73	55.88	85.54	73.41	66.34	79.83	61.39	81.56	80.22	79.30	68.69	75.17
Prompts-only	80.27	45.19	69.22	32.24	82.80	74.30	64.00	80.99	50.00	73.87	66.79	71.23	62.61	67.58
FedLMG	78.16	52.56	69.22	62.67	82.32	<u>76.46</u>	70.23	81.58	63.12	83.76	75.32	79.30	69.61	75.45
FedDEO	79.50	<u>53.99</u>	<u>69.86</u>	63.39	84.94	73.16	70.81	<u>89.18</u>	60.40	<u>85.22</u>	78.04	78.25	72.80	77.31
FedBiP	80.46	53.78	68.96	<u>70.18</u>	85.71	77.74	<u>72.81</u>	86.26	78.96	84.98	<u>79.13</u>	<u>80.35</u>	<u>73.56</u>	<u>80.54</u>
Espresso	81.99	58.08	71.39	80.18	88.33	71.96	75.32	89.47	<u>75.74</u>	85.84	83.49	82.11	76.29	82.16

perform knowledge distillation by training a generator. (3) **Diffusion-based OSFL methods**, including Prompts-only where the server uses text prompts (e.g., "a [domain] of a [class]") as conditions; FedLMG [48] where the server uses local models for guidance and train a global model by knowledge distillation; FedDEO [50] where clients upload learned descriptors; and FedBiP [3] where clients upload learned concepts and noisy latent noises.

Implementation details. For all diffusion-based OSFL methods, we use Stable Diffusion v1.5 as the pre-trained diffusion model, and generate $5\times$ the size of the local dataset. On the server, we use ResNet-18 as the global model, which is initialized with parameters pre-trained on ImageNet [5]. For Fens, DENSE and FedLMG, the architecture of local models is the same as the global one. For FedLMG, we utilize specific-teacher distillation for global model training. For FedDEO, FedBiP and Espresso, the number of epochs of local training is set to 100 for better convergence. For Espresso, we employ the pre-trained DINOv2 [20] as the feature extractor Φ , and identify intra-category patterns by hierarchical clustering. The number of intra-category clusters J is 4 and that of components of GMMs M is 4. The threshold of applying Z-Sampling T_z is 800. More details are provided in Appendix.

5.2. Comparison Results

As demonstrated in Tab. 1 and 2, Espresso achieves the best performance across the three datasets in terms of average accuracy. Overall, Espresso shows significant advantage over traditional FL methods. Although variant methods like FedProx, FPL, and FedSR improve FedAvg to cope with different types of data heterogeneity, their

Table 2. Performance comparison on PACS dataset under the 16-shot setting. All values are accuracy (%).

Methods	PACS				
	A	C	P	S	Avg.
<i>Ceiling</i>	86.25	84.30	94.24	79.07	85.96
FedAvg	71.68	76.38	87.08	42.03	69.29
FedProx	71.04	74.26	89.66	41.19	69.04
FPL	72.87	75.42	87.10	47.29	70.67
FedSR	71.68	76.39	91.05	44.66	70.95
DENSE	65.94	68.90	75.21	61.48	67.88
FedD3	63.43	68.28	72.20	53.29	64.30
Fens	70.87	68.60	90.86	65.51	73.96
Prompts-only	74.60	72.56	88.47	47.97	70.90
FedLMG	70.71	72.70	<u>91.65</u>	61.70	74.19
FedDEO	76.05	76.66	92.05	60.76	76.38
FedBiP	<u>77.35</u>	<u>83.03</u>	89.86	<u>66.19</u>	<u>79.11</u>
Espresso	77.51	83.73	90.46	69.75	80.36

performance can still be affected due to the inherent difficulty of parameter aggregation under high data heterogeneity. Moreover, Espresso performs better than non-diffusion OSFL methods. We attribute this to the ability of Espresso to leverage the powerful generative prior of the diffusion model, which allows for better reconstruction of client data distributions. Moreover, compared to other diffusion-based OSFL, Espresso still shows an edge.

Notably, Espresso achieves best performance on some domains with significant stylistic divergence, such as

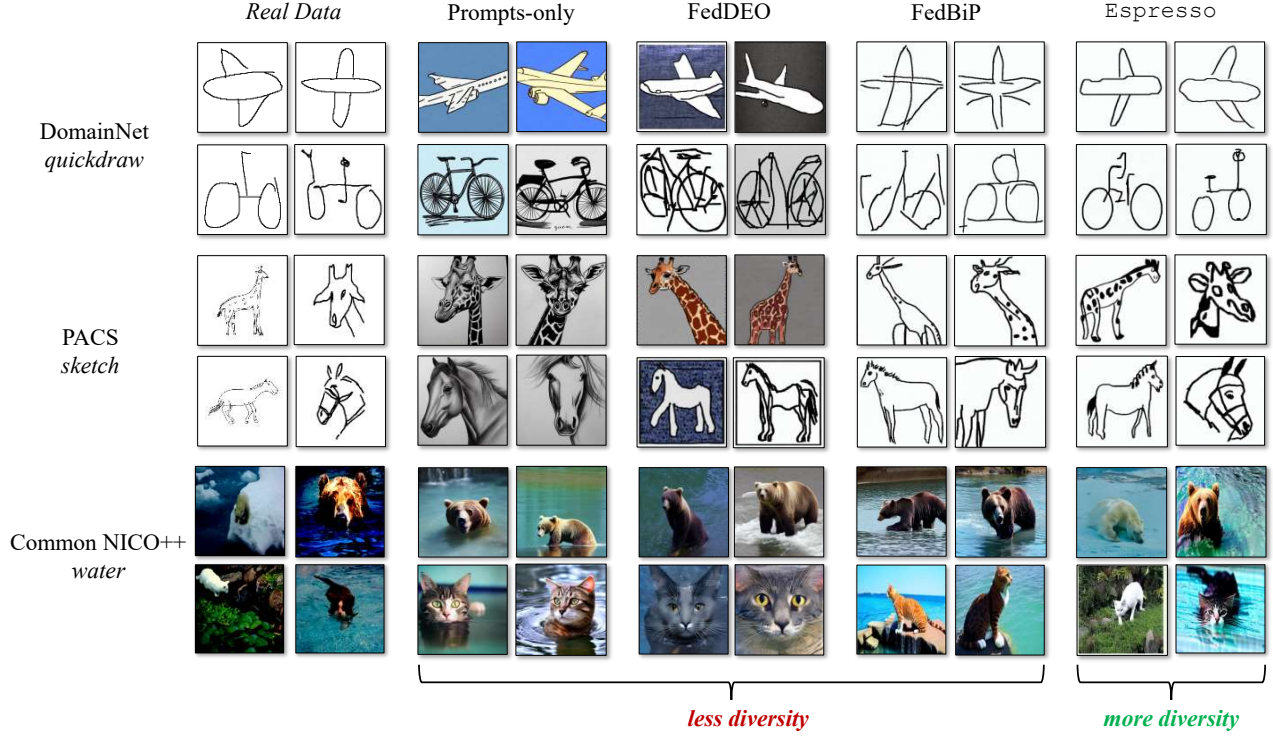


Figure 2. Visualization of synthetic data generated across datasets. Each row corresponds to a category. For each category, we show two images for each method. *Espresso* achieves better fidelity and diversity, and better reproduces different intra-category patterns.

the *quickdraw* in DomainNet, where it achieves an accuracy of 80.18%, demonstrating its effectiveness in capturing such unique and abstract styles. However, we also observe that its performance is slightly surpassed by competing methods in a few specific domains, such as the *sketch* in DomainNet (won by FedBiP) and the *photo* in PACS (won by FedDEO). We hypothesize this is because the pre-trained diffusion model already possesses a strong prior for these common styles, allowing simpler guidance mechanisms to also be effective. Nevertheless, *Espresso*'s performance across all datasets validates the importance of generating data with both high semantic fidelity and diversity. The comparison results for label-skew settings are in Appendix.

5.3. Visualizations

5.3.1. Visualizations of Synthetic Data

Fig. 2 shows synthetic images from *Espresso* and key baselines. While generating realistic images, *Prompts-only* fails to match target styles and struggles to reflect crucial intra-category patterns (e.g., a polar bear or a brown bear), which limits its usability in global model training. *FedDEO* shows similar limitations in capturing target styles and fine-grained details, likely due to its partial reliance on textual guidance during the process of denoising. *FedBiP* can reproduce target styles but usually compromises the object's structural integrity and also inadequately ren-

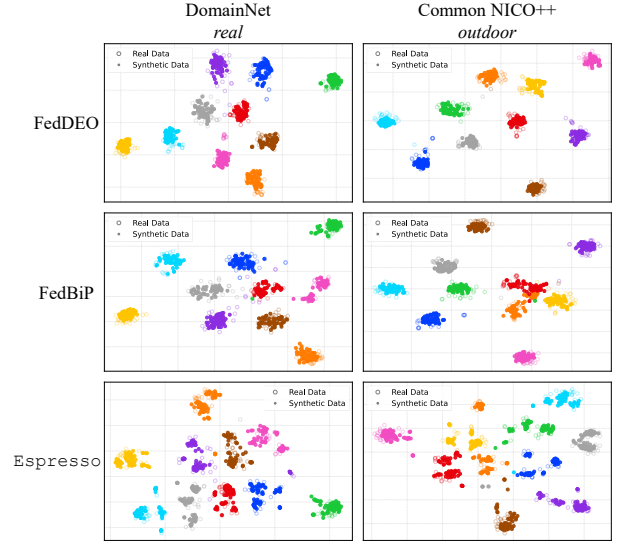


Figure 3. Visualization of synthetic images generated across diverse datasets. Each row corresponds to a method, with the top row displaying real data as ground truth.

ders intra-category differences. In contrast, *Espresso* can more accurately capture the target styles and generate structurally complete objects in these styles. In addition, it can also recognize and reproduces different intra-category pat-

terns in the synthetic data. This visual evidence supports the effectiveness of Espresso’s fine-grained condition learning on better reconstructing the distribution of local data.

5.3.2. Visualizations of Semantic Distribution

In order to analyze the semantic distribution of data generated by different methods, Fig. 3 shows the t-SNE visualization of the real and synthetic data of Espresso, FedDEO, and FedBiP on *real* in DomainNet and *outdoor* in Common NICO++. The features are extracted by DINOv2 [20]. Although FedDEO and FedBiP have high semantic fidelity, they form obvious but overly compact clusters, indicating the loss of intra category details. In contrast, Espresso is able to discover small patterns within each category and spread the synthetic data across these small clusters, making the distribution of synthetic data closer to the real data. This intuitively demonstrates that the fine-grained guidance mechanism of Espresso can more accurately capture intra-category patterns missed by other methods.

5.4. Ablation Study

To evaluate the impact of each component, we conduct an ablation study with the results shown in Fig. 4. The Prompts-only baseline consistently performs the worst, confirming that client-specific guidance is essential. Enabling Fine-Grained Condition Learning (C) alone provides the most substantial performance gain, establishing that it is effective to capture high semantic fidelity and diversity for better performance. The other components, GMM of Latent Space (G) and Z-Sampling (Z), primarily demonstrate their effectiveness through their synergistic effect with C, which is more evident on PACS. Ultimately, the complete Espresso (C+G+Z) achieved the highest accuracy across all datasets, indicating that when all three components are complete and work together, they can collectively enhance the semantic fidelity and diversity of the synthetic dataset, and improve the performance of the global model.

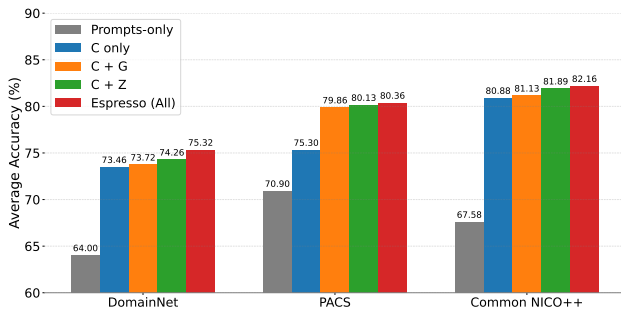


Figure 4. Ablation study for different component combinations: Fine-Grained Condition Learning (C), GMM of Latent Noises (G), and Z-Sampling (Z). The Prompts-only baseline represents data generation without any of our proposed components.

5.5. Parameter Analysis

To evaluate the sensitivity of the key hyperparameters of Espresso, we analyze the number of clusters J and GMM components M , as shown in Fig. 5. For different choices of J , which corresponds to the number of intra-category clusters, the performance on DomainNet and PACS consistently improves as the value increases. This validates our core hypothesis that modeling fine-grained intra-category patterns is helpful for improving the performance of the global model. Although performance on Common NICO++ peaks at 6, a value of J between 4 and 6 offers the best trade-off between performance and the number of conditional embeddings across all datasets. For different choices of M , the performance tends to plateau or fluctuate within a narrow range after reaching 2 to 4. This suggests that a moderately complex model is sufficient to model the distribution of the latent noises. It is worth noting that an overly simple GMM (1 component) still cannot capture enough distributional information, while an excessive number of components may create overly dense and low-variance components, thereby reducing the diversity of the initial latent noises and offering less performance benefit. More parameter analysis results are provided in Appendix.

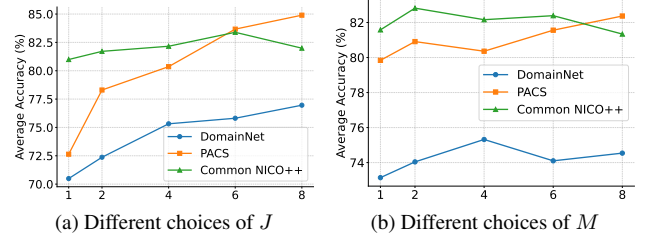


Figure 5. Parameter analysis of the hyperparameters.

6. Conclusions

In this paper, we address the challenges of insufficient semantic fidelity and diversity that constrain existing Diffusion-based One-shot Federated Learning (OSFL) methods. To overcome these limitations, we propose Espresso, a novel framework that enhances server-side data generation. Espresso utilizes Fine-Grained Condition Learning to capture fine-grained intra-category patterns for improved semantic fidelity and diversity, and employs Semantics-Preserving Sampling to enhance the alignment between synthetic data and learned conditional embeddings. Our core contribution is a method that effectively compresses client data into highly representative parameters, which then guide a diffusion model to reconstruct the clients’ data distributions with high fidelity on the server for global model training. Extensive experiments demonstrate that Espresso outperforms state-of-the-art methods by generating a high-quality synthetic dataset, thereby improving the performance of the global model.

Acknowledgements

The research is supported by the National Key R&D Program of China (2023YFB2703700), the National Natural Science Foundation of China (62176269), GMCC-SYSU Joint Lab for Smart Applications, Zhuhai Key Laboratory of Trusted Large Language Models.

References

- [1] Youssef Allouah, Akash Dhasade, Rachid Guerraoui, Nirupam Gupta, Anne-Marie Kermarrec, Rafael Pinot, Rafael Pires, and Rishi Sharma. Revisiting ensembling in one-shot federated learning. *Advances in Neural Information Processing Systems*, 37:68500–68527, 2024. [1](#), [2](#), [5](#)
- [2] Chuan Chen, Tianchi Liao, Xiaojun Deng, Zihou Wu, Sheng Huang, and Zibin Zheng. Advances in robust federated learning: A survey with heterogeneity considerations. *IEEE Transactions on Big Data*, 2025. [1](#), [3](#)
- [3] Haokun Chen, Hang Li, Yao Zhang, Jinhe Bi, Gengyuan Zhang, Yueqi Zhang, Philip Torr, Jindong Gu, Denis Krompass, and Volker Tresp. Fedbip: Heterogeneous one-shot federated learning with personalized latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30440–30450, 2025. [1](#), [2](#), [3](#), [5](#), [6](#)
- [4] 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*. [2](#)
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [6](#)
- [6] Yiqun Diao, Qinbin Li, and Bingsheng He. Towards addressing label skews in one-shot federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. [1](#), [2](#)
- [7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. [2](#)
- [8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*. [2](#)
- [9] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyang Wu, Terrence Chen, David Doermann, and Arun Innanje. Preserving privacy in federated learning with ensemble cross-domain knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11891–11899, 2022. [1](#), [2](#)
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. [1](#)
- [11] Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019. [1](#)
- [12] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. [4](#)
- [13] Clare Elizabeth Heinbaugh, Emilio Luz-Ricca, and Hua-jie Shao. Data-free one-shot federated learning under very high statistical heterogeneity. In *The Eleventh International Conference on Learning Representations*, 2023. [2](#)
- [14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. [1](#)
- [15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [3](#)
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1](#), [2](#), [3](#)
- [17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. [2](#)
- [18] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. [2](#), [5](#)
- [19] Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Towards a theoretical and practical understanding of one-shot federated learning with fisher information. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023. [2](#)
- [20] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. [4](#), [6](#), [8](#)
- [21] Anirudh Kasturi, Anish Reddy Ellore, and Chittaranjan Hota. Fusion learning: A one shot federated learning. In *International Conference on Computational Science*, pages 424–436. Springer, 2020. [2](#)
- [22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. [5](#)
- [23] Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1484–1490. International Joint Conferences on Artificial Intelligence Organization, 2021. Main Track. [2](#)

- [24] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10713–10722, 2021. 1
- [25] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 35(4):3347–3366, 2021. 1, 3
- [26] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems, 2:429–450, 2020. 1, 2, 3, 5
- [27] Yichen Li, Wenchao Xu, Haozhao Wang, Yining Qi, Jingcai Guo, and Ruixuan Li. Personalized federated domain-incremental learning based on adaptive knowledge matching. In European conference on computer vision, pages 127–144. Springer, 2024. 1
- [28] Bai LiChen, Shitong Shao, Zipeng Qi, Haoyi Xiong, Zeke Xie, et al. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In The Thirteenth International Conference on Learning Representations. 2, 4, 5
- [29] Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James T Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. ACM Computing Surveys, 2024. 2
- [30] Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. Fedlpa: One-shot federated learning with layer-wise posterior aggregation. Advances in Neural Information Processing Systems, 37:81510–81548, 2024. 2
- [31] Xiang Liu, Zhenheng Tang, Xia Li, Yijun Song, Sijie Ji, Zemin Liu, Bo Han, Linshan Jiang, and Jialin Li. One-shot federated learning methods: a practical guide. In Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pages 10573–10581, 2025. 1
- [32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017. 1, 3, 5
- [33] A Tuan Nguyen, Philip Torr, and Ser Nam Lim. FedSr: A simple and effective domain generalization method for federated learning. Advances in Neural Information Processing Systems, 35:38831–38843, 2022. 2, 5
- [34] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1406–1415, 2019. 5
- [35] Xin Qi, Meixuan Li, Sijin Zhou, Wei Feng, and Zhuang Qi. Federated learning for science: A survey on the path to a trustworthy collaboration ecosystem. Authorea Preprints, 2025. 1
- [36] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In Proceedings of the 31st ACM international conference on multimedia, pages 3099–3107, 2023. 1
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021. 2, 4
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022. 1, 2
- [39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22500–22510, 2023. 2
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In International Conference on Learning Representations. 2, 3
- [41] 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), pages 1–10. IEEE, 2023. 1, 2, 5
- [42] Zhenheng Tang, Yonggang Zhang, Peijie Dong, Yiu-ming Cheung, Amelie Zhou, Bo Han, and Xiaowen Chu. Fusefl: One-shot federated learning through the lens of causality with progressive model fusion. Advances in Neural Information Processing Systems, 37:28393–28429, 2024. 2
- [43] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. arXiv preprint arXiv:2002.06440, 2020. 2
- [44] Naibo Wang, Wenjie Feng, Moming Duan, Fusheng Liu, See-Kiong Ng, et al. Data-free diversity-based ensemble selection for one-shot federated learning. Transactions on Machine Learning Research, 2023. 2
- [45] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. arXiv preprint arXiv:1811.10959, 2018. 1, 2
- [46] Jiahao Wu, Ming Hu, Yanxin Yang, Xiaofei Xie, ZeKai Chen, Chenyu Song, and Mingsong Chen. Rising from ashes: Generalized federated learning via dynamic parameter reset. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025. 1
- [47] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys, 56(4): 1–39, 2023. 1
- [48] Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. One-shot heterogeneous federated learning with

- local model-guided diffusion models. In Forty-second International Conference on Machine Learning. 1, 6
- [49] Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. Exploring one-shot semi-supervised federated learning with pre-trained diffusion models. In Proceedings of the AAAI conference on artificial intelligence, pages 16325–16333, 2024. 2
- [50] Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. Feddeo: Description-enhanced one-shot federated learning with diffusion models. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 6666–6675, 2024. 1, 2, 3, 6
- [51] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023. 2
- [52] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. IEEE transactions on pattern analysis and machine intelligence, 46(1):150–170, 2023. 2
- [53] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazani. Bayesian nonparametric federated learning of neural networks. In International conference on machine learning, pages 7252–7261. PMLR, 2019. 2
- [54] Hui Zeng, Minrui Xu, Tongqing Zhou, Xinyi Wu, Jiawen Kang, Zhiping Cai, and Dusit Niyato. One-shot-but-not-degraded federated learning. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 11070–11079, 2024. 2
- [55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017. 5
- [56] 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, 2022. 1, 2, 5
- [57] 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. 2
- [58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023. 2
- [59] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyang Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16036–16047, 2023. 5
- [60] Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. arXiv preprint arXiv:2009.07999, 2020. 1, 2