PROMPT-GUIDED ONE-SHOT FEDERATED LEARNING VIA SYNTHETIC DATA GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional Federated Learning (FL) requires the server to continuously establish wireless connections with clients to communicate learning models, which burdens the network in terms of management and wireless resource constraints, and may facilitate inference attacks. One-shot FL has emerged as a key innovation to reduce the framework's dependence on a consistent wireless connection. However, existing approaches largely treat clients as passive contributors: the algorithm determines how their local data are used, with limited support for heterogeneous user preferences. In this paper, we propose a one-shot FL framework that allows clients to actively control how their local data contribute to the learning process through synthetic data generation. Each client specifies its preference through a natural-language prompt, which guides generation together with its local images and controls how much the synthetic data deviate from the originals. In addition, we address the limited computational capability of client devices by incorporating a latent consistency model-enabled low-rank adaptation to boost the parameter and energy efficiency. Finally, we conduct a series of experiments to evaluate the effectiveness of our proposal against state-of-the-art benchmarks and obtain substantial accuracy improvements across different deep learning models and levels of heterogeneity across three datasets.

1 INTRODUCTION

Federated learning (FL) has enabled the high performance of deep learning models by leveraging the local data at clients to collaboratively train a global model McMahan et al. (2017). It has been adopted in a wide range of applications spanning many fields: smart farming Abu-Khadrah et al. (2023), medical analysis Pfitzner et al. (2021), and recently, large language models (LLMs) Ye et al. (2024). Despite the framework's advancement, the continuous downlink transmission of the global model and the uplink transmission of the local model both entail disadvantages. The frequent communication of learning models incurs high communication costs, model degradation due to noise in the wireless channel, and inference attacks. To address this, one-shot federated learning (OSFL) has emerged as a key solution that eliminates the iterative communication rounds by restricting training to a single round of interaction between clients and the server, thereby reducing exposure to noise and attacks.

Current work on OSFL can effectively reduce the communication complexity in two directions: 1) Model knowledge-based OSFL Guha et al. (2018); Li et al. (2021); Zhang et al. (2022); Dai et al. (2024) and 2) Synthesis data-based OSFL Xiong et al. (2023); Yang et al. (2024); Chen et al. (2025). Model knowledge-based OSFL proposes that the client trains local models to completion, after which the server aggregates these models or distills its knowledge through ensemble models. The approach typically involves an additional public dataset, which is impractical in real-world scenarios. On the other hand, in the data-based OSFL, each client locally synthesizes a compact representation of its data (e.g., distilled/synthetic samples, feature statistics, or diffusion-based generations) and uploads only those synthesized data/representations to the server. The server then trains the global model from the constructed pooled dataset. However, this paradigm suffers from a trade-off between communication and privacy guarantees, which depends on the quality of the local synthesis process.

Despite success in reducing communication costs and mitigating network attack threats, the current OSFL still faces major drawbacks. For instance, the dependency on the public dataset of the model-based approach, which is highly unrealistic in sensitive fields such as medical and surveillance. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

work Zhang et al. (2022) omitted this dependency by proposing an image synthesis generator at the server to distill knowledge from local models. On the contrary, the data-based approach can compromise the privacy of real images through the generation of synthetic samples. Both approaches suffer from severe performance degradation under increasing data heterogeneity and fail to explicitly account for heterogeneous, user-specific sharing preferences.

Motivated to address this gap, we propose a novel data-based OSFL that gives each client an active role in the OSFL framework while reducing the heterogeneity effects of client data on the performance of the global model. Specifically, we present a pre-trained vision-language model (VLM) at the client for image synthesis. The proposed VLM incorporates each client's preference as context prompts to guide the image synthesis process. The contributions of the paper are presented as follows:

- To accommodate heterogeneous user preferences, we make local data synthesis controllable by each client. Each client expresses its preference through a natural-language prompt that guides image generation. The client can further control how much the synthetic images differ from its original data by adjusting the image-conditioning scale and the number of denoising steps, and shares only the generated samples it approves.
- We integrate an image-prompt adapter that combines the client's preference prompt with its original image through decoupled cross-attention, with an image-conditioning scale controlling how closely the generated image follows the original. To support resource-constrained devices, we further incorporate a latent consistency model with low-rank adaptation (LCM-LoRA). This design reduces both per-step computation and the number of denoising steps.
- Finally, extensive experiments show that our framework outperforms state-of-the-art OSFL methods across different model architectures and levels of data heterogeneity. In addition, we provide synthetic examples reflecting different user preferences, showing that the generated data follow each client's specified preference.

2 Related Works

2.1 Model Knowledge-assisted OSFL

Guha et al. (2018) proposed unlabeled data to mimic the teacher model's output and then train the student model, which can significantly reduce the communication costs. Later, Li et al. (2021) proposed a two-stage knowledge transfer (client and server levels) framework that can be extended to any classification models without restriction. To eliminate the training strategy's dependence on the auxiliary public dataset, Zhang et al. (2022) leveraged ensemble models to train an image generator and then used the generated images to distill the ensemble's knowledge into the global model. Authors in FuseFL Tang et al. (2024) decompose the learning models into blocks, where they progressively share these blocks across clients to fuse and train them in a bottom-up manner with the objective to eliminate spurious features of local data. FedBens Talpini et al. (2025) proposed a mixture of Laplace approximation to estimate the client's local posteriors, and leverage the information to obtain a multimodal global posterior.

2.2 Data-assisted OSFL

The authors in Zhou et al. (2020) were the first to consider dataset distillation in the OSFL. Specifically, each client distills their private dataset and then sends the resulting synthetic data to the server for global training. Inspired by the work, Xiong et al. (2023) adopted the iterative distribution matching Zhao & Bilen (2023) approach for dataset condensation, resulting in richer information and capturing the global viewpoint. Wang et al. (2024) recognized the possibility of the client drift problem, where each client independently condensed the data. The authors proposed collaborative data condensation learning among clients, using the sliced Wasserstein distance as a regularization term to align local distribution with the broader distribution across other clients. On the other hand, Song et al. (2023a) addressed the resource-constrained problem for edge devices by distilling the dataset using a kernel ridge regression loss. Zhang et al. (2024) leverages a pre-trained auto-encoder model to assist the client in encompassing local information and $\mathcal{V}$-information-based Core-Set to extract a diverse and informative subset from local data. Yao et al. (2025) addresses diverse computing in OSFL by dividing clients into two groups: resource-sufficient and resource-constrained, employing

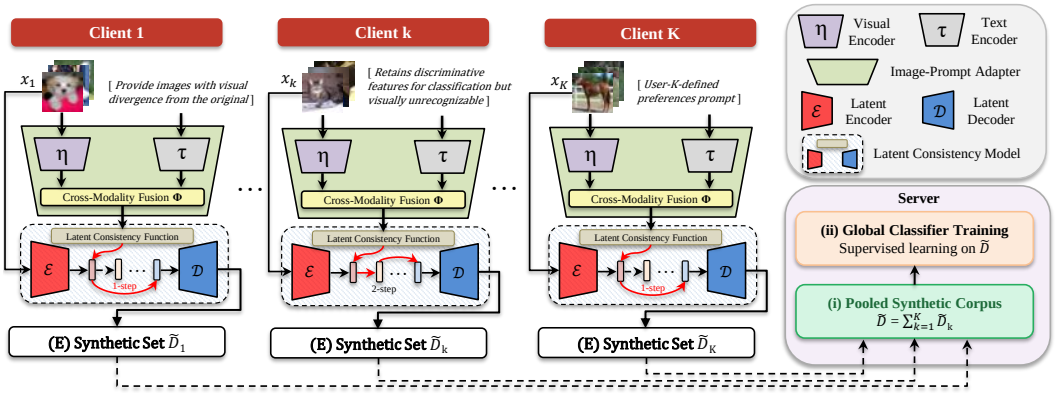

Figure 1: Overview of the proposed framework: clients steer local synthetic data generation with their own preference prompts, and the server trains the global classifier on the pooled synthetic data.

DL-based classification training for the former and lightweight generative models for the latter. Yan et al. (2025) proposed multi-round data condensation for medical images, where the condensation at each round targets and condenses knowledge that the current model does not contain. Therefore, they avoid unnecessary repetition of homogeneous knowledge and reduce the communication cost.

3 Proposed Framework

3.1 Preliminary

We consider a federated learning system that consists of a central server and K distributed clients. Client $k \in \{1, \ldots, K\}$ holds a private local dataset $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}, x_{k,i} \in \mathbb{R}^{C \times H \times W}$, $y_{k,i} \in \{1, \ldots, C\}$. The union of all private datasets forms the global private corpus $\mathcal{D} = \bigcup_{k=1}^{K} \mathcal{D}_k$. In conventional FL, these data are utilized to iteratively train the global model over a large number of communication rounds, incurring significant communication overhead and requiring persistent, reliable network connectivity due to frequent model exchanges between clients and the server. In addition, despite the FL framework's design to prevent exposure of client data, Geiping et al. (2020) has demonstrated the possibility of reconstructing high-resolution images using parameter gradients, which is extremely severe in the medical field. These drawbacks motivate us to redesign FL so that what each client shares is human-interpretable and controlled by the client itself. In our framework, each client shapes its synthetic contribution through a preference prompt and a small set of generation settings. This process requires only one-way communication from clients to the server.

3.2 Proposed VLM-assisted Client for Synthetic Data

We propose a vision-language generative model that synthesizes preference-guided images conditioned on the client-defined textual prompt, as shown in Figure 1. The textual prompt encodes the client's preferences, such as semantic or structural deviations from original images. To integrate this conditioning effectively, we propose an image-prompt adapter at the client to learn a joint representation between the image features and prompt embeddings, which is subsequently injected into the generative backbone. This enables fine-grained, client-controllable synthesis that balances semantic fidelity and deviation from the source. Given a local image $x_{k,i}$ and the client's prompt p_k, we first encode the prompt and image into semantic and visual embeddings:

$$\mathbf{c}_t = \tau(p_k), \quad \mathbf{c}_v = \eta(x_{k,i}), \tag{1}$$

where $\tau(\cdot)$ and $\eta(\cdot)$ are respectively the text and visual encoders of the IP-adapter. These embeddings are then fused into a unified conditioning representation with the decoupled cross-attention layers:

$$\mathbf{C} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} + \gamma \cdot \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}'^\top}{\sqrt{d}}\right)\mathbf{V}', \tag{2}$$

here the first branch captures text-conditioned guidance, the second branch captures image-conditioned guidance, and the scaling factor γ controls the influence of image-based conditioning.

Specifically, the query, key, and value matrices are defined as:

$$\mathbf{Q} = \mathbf{z}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{c}_t\mathbf{W}_k, \quad \mathbf{V} = \mathbf{c}_t\mathbf{W}_v, \quad \mathbf{K}' = \mathbf{c}_v\mathbf{W}'_k, \quad \mathbf{V}' = \mathbf{c}_v\mathbf{W}'_v, \tag{3}$$

where $\mathbf{z}$ is the query feature. Our design enables the model to jointly enforce semantic constraints from the prompt and structural priors from the input image, allowing flexible control over the deviation-utility trade-off. The constructed joint conditioning signal is then fed to the diffusion model to produce synthetic output:

$$\tilde{x}_{k,i} = G_\theta(x_{k,i}, N, \mathbf{C}), \tag{4}$$

where N is the number of denoising steps. A standard diffusion model process gradually corrupts a clean image x_0 into a noisy sample through a Markovian forward process: $q(x_n \mid x_{n-1}) = \mathcal{N}\big(x_n; \sqrt{1-\beta_n}\, x_{n-1},\, \beta_n \mathbf{I}\big)\,, n = 1, \ldots, N$, where $\beta_n \in (0, 1)$ is the variance schedule and N denotes the total number of diffusion steps Sohl-Dickstein et al. (2015). The corresponding reverse process learns to denoise the sample step by step, conditioned on the joint representation $\mathbf{C}$:

$$p_\theta(x_{n-1} \mid x_n, \mathbf{C}) = \mathcal{N}(x_{n-1}; \mu_\theta(x_n, n, \mathbf{C}), \Sigma_\theta(x_n, n))\,, \tag{5}$$

where the model iteratively transforms a Gaussian noise sample $x_n \sim \mathcal{N}(0, \mathbf{I})$ into a synthesized image x_0. Regardless of the high quality of the generated images, conventional diffusion Ho et al. (2020) inference remains computationally expensive for resource-constrained clients for two main reasons: pixel-level denoising and the large number of iterative sampling steps N.

To address this, we incorporate a latent consistency model to improve efficiency. Specifically, the latent model operates in a compressed space $z \in \mathbb{R}^{c \times h \times w}$, where $c \times h \times w \ll C \times H \times W$, as determined by the downsampling factor. This significantly reduces the computational cost of each denoising step. The process can be formulated as

$$z_N = \mathcal{E}(x_{k,i}), \quad z_{n-1} = f_\theta(z_n, n, \mathbf{C}), \quad \tilde{x}_{k,i} = \mathcal{D}(z_0), \tag{6}$$

where $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ are the pre-trained variational auto-encoder/decoder Rombach et al. (2022). $f_\theta(\cdot)$ represents the latent diffusion model that iteratively denoises the latent variable conditioned on $\mathbf{C}$. On the other hand, the self-consistency of the latent consistency model enables it to directly predict the origins of the augmented probability flow ordinary differential equation from any points on the trajectory, thereby eliminating many iterations Song et al. (2023b); Luo et al. (2023a).

$$z_0 \approx F_\theta(z_t, t, \mathbf{C}) \approx F_\theta(z_{t'}, t', \mathbf{C}), \quad \forall\, t, t' \in \{1, \ldots, N\}. \tag{7}$$

This property facilitates the model to skip the sequence of inference steps. Furthermore, a low-rank adaptation technique Hu et al. (2022) is considered to support the fine-tuning of the LCM, thus improving parameter efficiency Luo et al. (2023b). The image synthesis process is now presented as:

$$\tilde{x}_{k,i} = \mathcal{D}\big(F_{\theta+\Delta\theta_{LoRA}}(\mathcal{E}(x_{k,i}), N', \mathbf{C})\big), \tag{8}$$

where $F_{\theta+\Delta\theta_{LoRA}}$ is the LCM-based direct latent mapping with the assistance from the LoRA mechanism to achieve the parameter-efficient fine-tuning for the VLM, inference step $N' \ll N$, and its typical values are from 2 to 5 steps. It is worth noticing that not all generated synthesis images have to be delivered to the central server; the client can double-check the synthesis image before transmitting it to the server. We later reflect this property by considering the fraction of synthesis data that can be delivered to the server. Each client uploads $\widetilde{\mathcal{D}}_k$ exactly once, no local optimization traces are communicated, and no repeated server-client synchronization is required.

3.3 Model Training with Synthesis Data at Server

After collecting all accepted synthetic datasets, the server forms the pooled synthetic corpus $\widetilde{\mathcal{D}} = \bigcup_{k=1}^{K} \widetilde{\mathcal{D}}_k$ and trains the global classifier by empirical risk minimization:

$$\phi^\star = \arg\min_{\phi} \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{(\tilde{x}, \tilde{y}) \in \widetilde{\mathcal{D}}} \ell_{\mathrm{CE}}(f_\phi(\tilde{x}), \tilde{y}), \tag{9}$$

where $\ell_{\mathrm{CE}}(\cdot)$ is the cross-entropy loss. In addition to the empirical results shown in Section 4, we conduct a convergence analysis of the classifier on synthetic data, and then indicate its optimum is close to the real-data optimum, as shown in Supplementary A. The pipeline contains only a single communication stage from clients to the server, followed by centralized training.

Table 1: Accuracy results under heterogeneous levels α and are grouped by backbone architecture.

Backbone	Method	CIFAR10			SVHN			CIFAR100		
		0.05	0.1	0.5	0.05	0.1	0.5	0.05	0.1	0.5
CNN	FedAvg (One-shot)	$39.89_{\pm 3.1}$	$43.98_{\pm 3.8}$	$55.98_{\pm 1.4}$	$46.58_{\pm 3.0}$	$59.93_{\pm 6.1}$	$81.23_{\pm 1.0}$	$8.31_{\pm 1.5}$	$12.63_{\pm 0.6}$	$28.30_{\pm 0.6}$
	DENSE 2022	$23.37_{\pm 6.6}$	$39.98_{\pm 7.1}$	$63.25_{\pm 3.6}$	$38.72_{\pm 6.9}$	$52.91_{\pm 5.9}$	$74.57_{\pm 2.4}$	$18.53_{\pm 1.6}$	$23.18_{\pm 0.8}$	$36.08_{\pm 0.6}$
	FedDM 2023	$47.28_{\pm 1.4}$	$49.57_{\pm 0.6}$	$57.2_{\pm 1.5}$	$62.69_{\pm 3.3}$	$72.12_{\pm 2.5}$	$80.92_{\pm 0.3}$	$26.45_{\pm 0.2}$	$28.45_{\pm 0.4}$	$28.45_{\pm 0.4}$
	FedCGS 2025	$59.52_{\pm 0.3}$	$59.51_{\pm 0.3}$	$59.52_{\pm 0.3}$	$69.21_{\pm 0.6}$	$69.17_{\pm 0.6}$	$69.20_{\pm 0.6}$	$31.85_{\pm 0.4}$	$31.86_{\pm 0.4}$	$31.86_{\pm 0.4}$
	Ours (Diversity)	$69.98_{\pm 0.7}$	$\mathbf{72.10}_{\pm 0.5}$	$\mathbf{72.58}_{\pm 0.6}$	$74.28_{\pm 4.5}$	$76.75_{\pm 4.3}$	$84.69_{\pm 0.6}$	$38.22_{\pm 0.8}$	$40.51_{\pm 0.8}$	$44.29_{\pm 0.5}$
	Ours (Consensus)	$\mathbf{70.53}_{\pm 0.2}$	$71.15_{\pm 0.4}$	$70.86_{\pm 0.2}$	$\mathbf{85.46}_{\pm 0.2}$	$\mathbf{85.63}_{\pm 0.1}$	$\mathbf{85.55}_{\pm 0.2}$	$\mathbf{55.30}_{\pm 0.3}$	$\mathbf{55.39}_{\pm 0.1}$	$\mathbf{55.88}_{\pm 0.2}$
ResNet-18	FedPFT 2024	$56.39_{\pm 0.1}$	$56.57_{\pm 0.1}$	$57.45_{\pm 0.3}$	$42.72_{\pm 0.5}$	$43.26_{\pm 0.2}$	$43.42_{\pm 0.4}$	$36.85_{\pm 0.3}$	$37.43_{\pm 0.1}$	$38.52_{\pm 0.4}$
	Co-Boosting 2024	$44.37_{\pm 0.4}$	$60.41_{\pm 0.7}$	$67.43_{\pm 0.4}$	$41.90_{\pm 0.4}$	$57.13_{\pm 0.3}$	$84.65_{\pm 0.2}$	$20.30_{\pm 0.8}$	$24.63_{\pm 0.6}$	$34.43_{\pm 0.5}$
	FedCGS 2025	$63.98_{\pm 0.0}$	$63.99_{\pm 0.0}$	$63.99_{\pm 0.0}$	$57.79_{\pm 0.0}$	$57.79_{\pm 0.0}$	$57.79_{\pm 0.0}$	$39.96_{\pm 0.0}$	$39.99_{\pm 0.1}$	$39.96_{\pm 0.4}$
	FAFI 2025	$61.61_{\pm 5.9}$	$69.95_{\pm 4.0}$	$76.07_{\pm 0.8}$	$63.78_{\pm 1.9}$	$71.26_{\pm 3.9}$	$77.20_{\pm 1.2}$	$35.35_{\pm 3.1}$	$40.36_{\pm 0.7}$	$50.37_{\pm 2.8}$
	Ours (Diversity)	$75.68_{\pm 3.3}$	$77.12_{\pm 0.9}$	$80.74_{\pm 1.6}$	$78.81_{\pm 4.3}$	$80.14_{\pm 4.7}$	$88.26_{\pm 0.8}$	$51.08_{\pm 2.0}$	$54.71_{\pm 2.5}$	$56.39_{\pm 1.8}$
	Ours (Consensus)	$\mathbf{78.10}_{\pm 1.2}$	$\mathbf{78.23}_{\pm 2.7}$	$\mathbf{80.86}_{\pm 1.2}$	$\mathbf{88.86}_{\pm 1.2}$	$\mathbf{88.91}_{\pm 0.8}$	$\mathbf{89.50}_{\pm 0.6}$	$\mathbf{59.54}_{\pm 0.9}$	$\mathbf{62.79}_{\pm 0.7}$	$\mathbf{63.53}_{\pm 0.3}$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets & Partitions: To evaluate the generalization capability of the proposed framework, we conduct extensive experiments on image classification tasks using three widely adopted federated learning benchmarks: CIFAR-10 Krizhevsky et al. (2009), SVHN Netzer et al. (2011), and CIFAR-100. For each dataset, the original training set is partitioned across clients to simulate realistic non-IID data distributions. Specifically, we employ a Dirichlet distribution with concentration parameter $\alpha \in \{0.05, 0.1, 0.5\}$ to control the degree of label heterogeneity among clients. Smaller values of α result in more skewed label distributions, where each client observes only a limited subset of classes, thereby creating highly heterogeneous data settings. Conversely, larger α values lead to more balanced data distributions across clients. After training, the global model is evaluated on the corresponding centralized test sets to assess its generalization performance under varying levels of data heterogeneity.

Baselines: To provide a fair comparison of our proposal against existing benchmarks, we evaluate it alongside several one-shot federated learning approaches. For consistency, we adapt FedAvg to a one-shot setting, where each client trains its local model to convergence before uploading it to the global server in a single communication round. Similarly, FedDM is configured such that clients perform a one-time distillation of their local datasets prior to transmission. Furthermore, we reproduce the performance of several state-of-the-art methods, including DENSE Zhang et al. (2022), FedPFT Peng et al. (2024), Co-Boosting Dai et al. (2024), FedCGS Guan et al. (2025), and FAFI Hui et al. (2025). This ensures a fair and meaningful evaluation aligned with the one-shot paradigm.

Configurations: Following prior work Dai et al. (2024); Guan et al. (2025), we consider ten clients, all of which participate in the one-shot FL process. Each client uses a lightweight synthesis pipeline based on `stable-diffusion-v1-5` Rombach et al. (2022). We integrate `lcm-lora-sdv1-5` Luo et al. (2023b) to reduce the number of denoising steps and `IP-Adapter` Ye et al. (2023) to condition generation on the client's local images. Together, these components enable efficient and controllable synthetic data generation on resource-constrained clients. For classification, we train the CNN from scratch, while ResNet-18 is initialized with ImageNet pre-trained weights for consistency with prior work.

4.2 EXPERIMENTAL RESULTS

We evaluate our framework under two representative scenarios: diversity and consensus. The diversity scenario captures heterogeneous client preferences by varying both the textual prompts and the control parameter γ. Specifically, clients may impose different generation objectives, such as "producing images that are visually distinct from the reference while preserving class identity," "removing background information to prevent leakage, resulting in isolated objects on a neutral backdrop," or "retaining discriminative features for classification while rendering the output visually unrecognizable from the original." In contrast, the consensus scenario assumes aligned client objectives, where

Table 2: Impact of the data sharing ratio λ on model performance under heterogeneity levels α.

Dataset	α	$\lambda = 0.1$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1.0$
CIFAR-10	0.05	$55.08_{\pm 1.2}$	$59.23_{\pm 0.4}$	$61.66_{\pm 0.6}$	$63.84_{\pm 0.4}$	$67.14_{\pm 0.5}$	$68.80_{\pm 0.2}$	$70.53_{\pm 0.2}$
	0.1	$55.86_{\pm 0.4}$	$59.64_{\pm 0.5}$	$61.43_{\pm 0.4}$	$64.39_{\pm 0.5}$	$67.60_{\pm 0.1}$	$69.16_{\pm 0.9}$	$71.15_{\pm 0.4}$
	0.5	$56.03_{\pm 0.4}$	$59.97_{\pm 0.3}$	$61.78_{\pm 0.4}$	$65.07_{\pm 0.8}$	$67.75_{\pm 1.0}$	$69.18_{\pm 0.2}$	$70.86_{\pm 0.2}$
SVHN	0.05	$77.79_{\pm 0.1}$	$79.42_{\pm 0.1}$	$81.30_{\pm 0.1}$	$82.14_{\pm 0.1}$	$83.65_{\pm 0.1}$	$84.32_{\pm 0.2}$	$85.46_{\pm 0.2}$
	0.1	$77.48_{\pm 0.8}$	$79.46_{\pm 0.4}$	$81.14_{\pm 0.3}$	$82.36_{\pm 0.4}$	$83.89_{\pm 0.1}$	$84.62_{\pm 0.2}$	$85.63_{\pm 0.1}$
	0.5	$77.64_{\pm 0.4}$	$79.49_{\pm 0.4}$	$80.87_{\pm 0.5}$	$82.43_{\pm 0.1}$	$83.82_{\pm 0.1}$	$84.83_{\pm 0.2}$	$85.55_{\pm 0.2}$
CIFAR-100	0.05	$29.31_{\pm 0.7}$	$35.46_{\pm 0.2}$	$39.25_{\pm 0.2}$	$43.32_{\pm 0.1}$	$48.87_{\pm 0.2}$	$52.90_{\pm 0.1}$	$55.30_{\pm 0.3}$
	0.1	$30.01_{\pm 0.8}$	$36.04_{\pm 0.3}$	$39.39_{\pm 0.4}$	$44.34_{\pm 0.3}$	$49.41_{\pm 0.1}$	$52.99_{\pm 0.2}$	$55.39_{\pm 0.1}$
	0.5	$30.23_{\pm 0.5}$	$35.67_{\pm 0.3}$	$39.31_{\pm 0.2}$	$43.92_{\pm 0.2}$	$48.97_{\pm 0.5}$	$53.25_{\pm 0.3}$	$55.88_{\pm 0.2}$

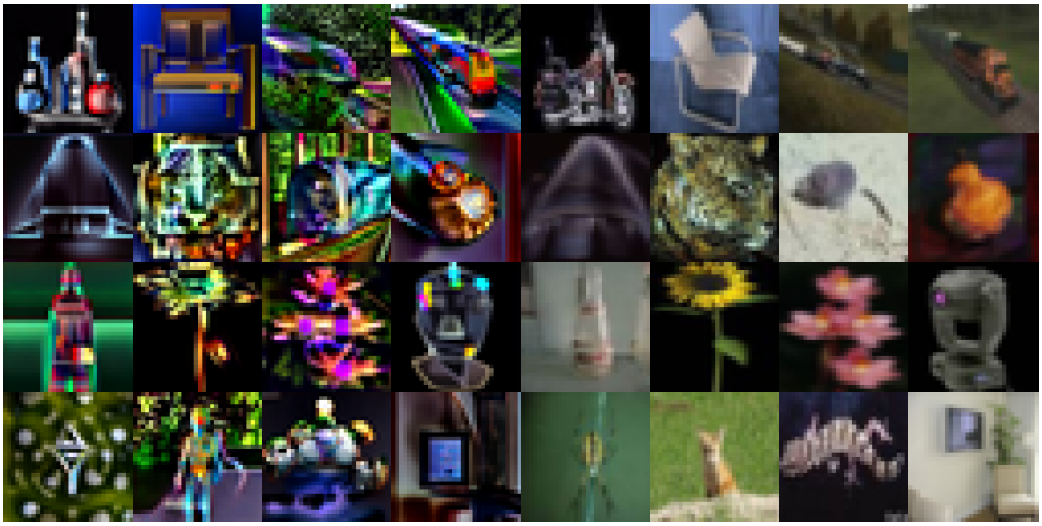

Figure 2: The visual comparison between the synthetic data and their top-1 similarity from the original dataset (CIFAR-100). The synthetic data are in the first four columns, and the last four columns contain the corresponding top-1 retrieval.

prompts and generation constraints are consistent across the network, reflecting a cooperative setting with shared preferences and utility requirements.

Accuracy Comparison: As shown in Table 1, the two schemes in our proposal outperform other SOTA OSFL baselines across all levels of label heterogeneity for all three datasets. The gains are especially significant under highly non-IID settings, demonstrating the robustness of our scenario to data heterogeneity. Under the CNN architecture, the accuracy gaps between the second best and our proposal (diversity) are around 10%, 4%, and 8% for CIFAR-10, SVHN, and CIFAR-100, respectively. On the other hand, with ResNet-18, the performance gap widens further, indicating that our approach scales effectively with stronger models and bootstraps efficiently.

Fraction of Data: As described in Section 3.2, following the image generation process with LCM-LoRA, clients can selectively filter shared data according to their own preferences. Table 2 analyzes system performance when only a fraction of the generated data (consensus setting) is shared with the server, where λ denotes the sharing ratio per client. Interestingly, sharing only 15% of the generated data already yields performance comparable to, or exceeding, most baselines, while increasing λ to 20% leads to consistent improvements across all datasets. We further observe that performance improves rapidly as λ increases from 10% to 50%; however, the improvement is slow beyond the 50% threshold. This suggests that transmitting the full dataset is unnecessary, as most performance gains are achieved with a limited subset. The phenomenon aligns with the Pareto principle, where a small portion of highly informative samples contributes disproportionately to the overall learning performance, while the remaining data introduces diminishing returns.

Preference Settings and Visualization: In Figure 2, we provide the synthetic data generated by two different prompts: the first prompt is "A $\{classname\}$ in a completely different visual style, different color palette, different texture, different angle, same semantic class but visually distinct from

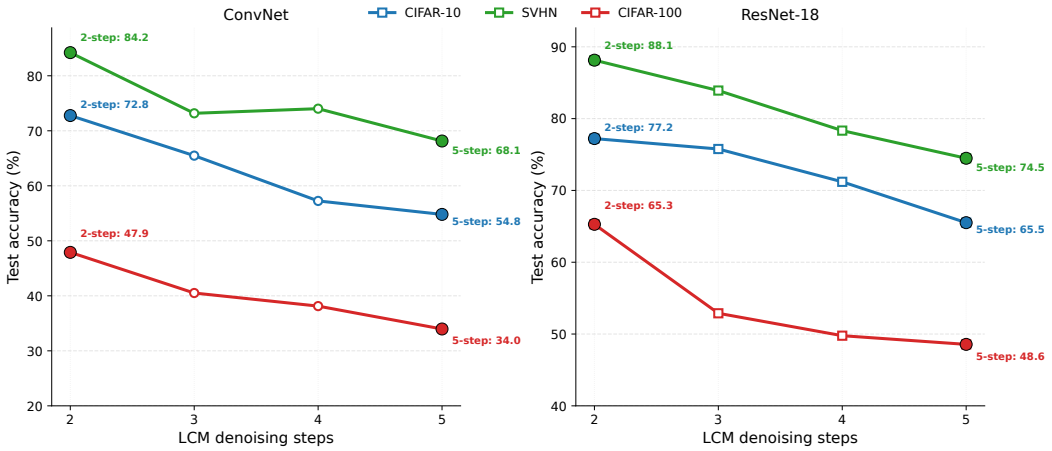

Figure 3: The accuracy of global models trained on generated images with different inference steps.

Table 3: Comparison of computing complexity in GFLOPs between the LCM-LoRA model and Stable Diffusion under different image resolutions.

Method	**Steps**	**32×32×3**	**56×56×3**	**112×112×3**	**224×224×3**
LCM-LoRA	2	12.52	25.76	77.56	269.10
	3	18.78	38.64	116.34	403.65
Stable Diffusion	20	250.60	515.00	1551.20	5382.00

any reference, creative reinterpretation" and "An abstract schematic of $\{classname\}$, preserving only essential shape silhouette and key structural features, heavily stylized, impossible to visually match to any real image, artistic geometric interpretation, unrecognizable appearance, useful for machine learning training". To make it easy to refer, the former prompt is called visual divergence and the latter is "feature preserving visual unrecognizable". Synthetic images from rows 1 and 2 are the results of the first prompt, while images from the remaining rows are from the latter prompt.

As we can see in the figure, the synthetic images exhibit properties that align closely with the client-defined prompts. Specifically, for the client's "visual divergence" prompt, the generated images preserve class-discriminative information while introducing noticeable variations in color, style, and texture compared to the original data. This indicates that the generation process successfully decouples semantic content from low-level visual attributes, enabling diversity without sacrificing task relevance. On the other hand, clients adopting the strong-abstraction prompt of "feature preserving and visually unrecognizable" produce images that reveal minimal visual cues about their original class and are difficult for human perception to interpret. Despite this, these images retain latent discriminative features that remain useful for downstream tasks, suggesting that the model captures abstract representations beyond human-recognizable patterns. This highlights the capability of our approach to balance deviation from the source and task utility under heterogeneous client preferences.

Inference Steps: Up to this point, we have demonstrated the performance of the system under a fixed number of denoising steps for the LCM process. In Figure 3, we present a comprehensive analysis of the relationship between the number of denoising steps at clients and model accuracy at the server. Overall, the accuracy of the learning model at the server decreases as the N' increases, which can be explained by the synthetic images. In particular, the denoising steps control the degree of change applied to the original images in response to the client context prompt, and with a large number of inference steps, the diffusion model can overshoot the client's preferences, degrading the quality of the generated images. Additionally, the cost of a single denoising step is 6.26 GFLOPs. On the other hand, with a few inference steps and a high value of γ, the synthetic images stay close to their originals and may retain content the client intended to change. Therefore, achieving a balanced state that respects the client's preferences and retains high semantic information is essential for the proposed system. Through a series of experiments, we have obtained a range of $\gamma \in \{0.65 - 0.8\}$ and the inference steps $N' \in \{2, 3, 4\}$, which provide a balanced sweet spot between deviation and accuracy.

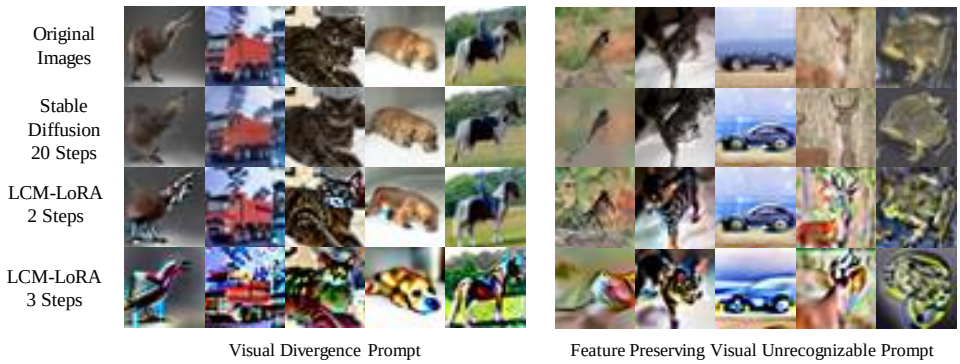

Figure 4: The efficiency of Latent Consistency Model-LoRA against the conventional Stable Diffusion Model in terms of the image synthesis process.

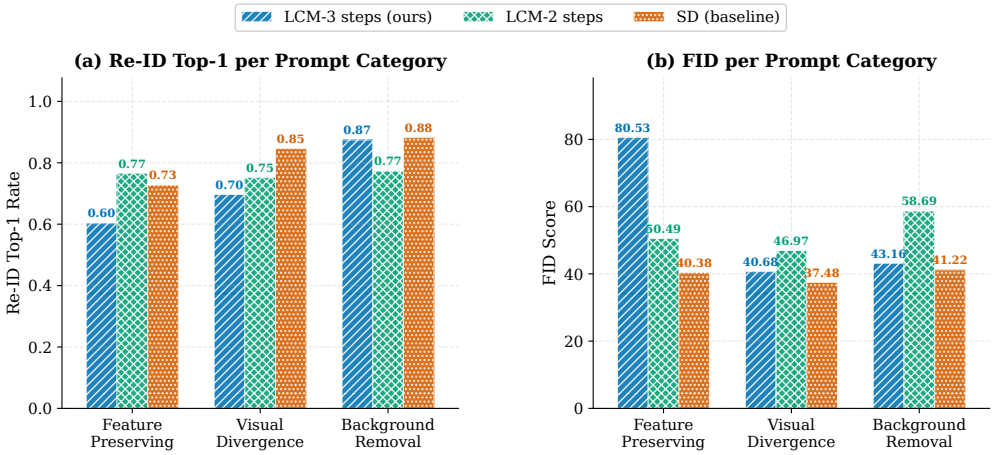

Figure 5: Comparison of Re-ID and FID between images generated by Latent Consistency Model-LoRA and the conventional Stable Diffusion Model across three prompt categories.

Computation Complexity: We provide the computation complexity of the current LCM-LoRA versus the original stable diffusion, as shown in Table 3. Our system requires significantly fewer inference steps compared to the conventional SD to adjust the original image according to the client's text prompt, approximately $6.67 \sim 10\times$ efficient. In addition, for each inference step, the proposed approach consumes only half as many computing resources as the SD. Therefore, the proposed LCM-LoRA achieves $13.34 \sim 20\times$ higher computing efficiency than the naive approach, thereby accelerating the image generation process.

In Figure 4, we provide the synthetic images from two approaches, along with the original images, to illustrate the generative capacity of the model under the client's prompt preferences. We observe that LCM-based generation exhibits stronger adherence to client-defined prompts compared to the conventional stable diffusion model. The consistency distillation process maps the latent representation directly to samples, thereby preventing iterative drift. On the other hand, multi-step diffusion progressively refines samples, which can deviate from prompt constraints.

Preference Adherence: Up to this point, we have assessed adherence to client preferences mainly from a visual perspective, which is inevitably influenced by human judgment. Figure 5 compares Re-ID Gheissari et al. (2006) and FID Heusel et al. (2017) across the three prompt categories. In the two appearance-altering settings, Feature Preserving and Visual Divergence, LCM-3 achieves lower Re-ID than SD, reducing the score from 0.73 to 0.60 in Feature Preserving and from 0.85 to 0.70 in Visual Divergence. This indicates that LCM makes the generated samples harder to trace back to the original training data while still following the client prompt. Background Removal is different, since the goal is only to change the background and keep the main object intact. In that case, a higher Re-ID is expected because the foreground semantics should remain similar. Even so, LCM remains

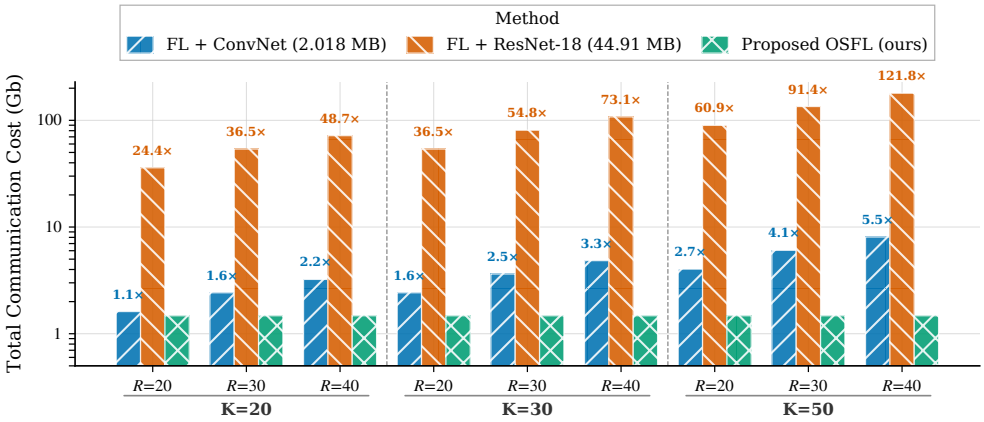

Figure 6: The communication efficiency of the proposed OSFL versus the conventional FL under a wide range of communication rounds and client number.

competitive, while requiring only 3 inference steps instead of 20 for SD. Overall, the results suggest that LCM better matches the intent of different prompts with much lower generation cost.

Communication Cost: In our proposed frameworks, we transmit the synthetic images from clients to the global server instead of learning models. Therefore, the communication cost depends on the data size and the number of data samples. Regardless of this, compared with conventional FL frameworks, our approach achieves communication efficiency as the number of clients scales, and is independent of the number of iterative communication rounds and model size. Figure 6 illustrates that the communication cost of our proposed framework remains consistent as the number of clients increases or the number of global rounds, and only depends on the number of data samples, we consider $\lambda = 0.3$. On the other hand, the communications of the conventional framework increase rapidly with the communication rounds and the size of learning models. For the ConvNet model with a size of 2.018 MB, we achieve $1.1\times$ communication efficiency with 20 communication rounds and 20 clients. This number goes up to $5.5\times$ when $K = 50$ and $R = 40$. This number for the ResNet-18 is much larger, ranging from 24.4 and up to 121.8.

5 Conclusion

In this paper, we present a novel one-shot federated learning framework that allows each client to explicitly define its preferences through a context prompt, which conditions the generation of synthetic data. By combining a pre-trained vision-language model with the original local data and a client-defined preference prompt, our framework enables clients to generate synthetic samples that preserve task-relevant semantics while reducing unwanted information leakage. To make the system practical for resource-constrained devices, we further incorporated a latent consistency model to accelerate generation by reducing the diffusion model's computational overhead through fewer inference steps. Extensive experiments on three benchmark datasets demonstrate that our approach consistently outperforms existing baselines across different levels of data heterogeneity, while offering a flexible and efficient solution for user-steerable one-shot federated learning.

AI use statement

In this work, we used generative AI tools solely to aid and polish the writing of the manuscript, including grammar correction, improving clarity and conciseness. We have not used generative AI tools for research ideation, methodology design, experimental design or execution, code development, or data analysis; the proposed method, all experiments, and all reported results are the work of the authors. Retrieval and literature search were likewise conducted by the authors. We have reviewed all AI-assisted writing: every suggested edit was checked by the authors for factual accuracy and consistency with the reported results, and no claim, number, or citation was adopted without verification against the underlying experiments and original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ahmed Abu-Khadrah, Ali Mohd Ali, and Muath Jarrah. An amendable multi-function control method using federated learning for smart sensors in agricultural production improvements. *ACM Transactions on Sensor Networks*, 2023.

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*, pp. 30440–30450, 2025.

Rong Dai, Yonggang Zhang, Ang Li, Tongliang Liu, Xun Yang, and Bo Han. Enhancing one-shot federated learning through data and ensemble co-boosting. In *The Twelfth International Conference on Learning Representations*, 2024.

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

N. Gheissari, T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1528–1535, 2006. doi: 10.1109/CVPR.2006.223.

Zenghao Guan, Yucan Zhou, and Xiaoyan Gu. Capture global feature statistics for one-shot federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16942–16950, 2025.

Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *2nd Workshop on Machine Learning on the Phone and other Consumer Devices, NeurIPs*, 2018.

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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Zeng Hui, Huang Wenke, Zhou Tongqing, Wu Xinyi, Wan Guancheng, Chen Yingwen, and Cai Zhiping. Does one-shot give the best shot? mitigating model inconsistency in one-shot federated learning. In *Proceedings of International Conference on Machine Learning (ICML)*, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis*, 2009.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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*, pp. 1484–1490, 2021.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *Preprint arXiv:2310.04378*, 2023a.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick Von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, pp. 4. Granada, 2011.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Z Peng, X Fan, Y Chen, Z Wang, S Pan, C Wen, R Zhang, and C Wang. Fedpft: Federated proxy fine-tuning of foundation models. In *Proceedings 33rd International Joint Conference on Artificial Intelligence*, 2024.

Bjarne Pfitzner, Nico Steckhan, and Bert Arnrich. Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, 21(2):1–31, 2021.

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*, pp. 10684–10695, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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 *Proceedings IEEE International Joint Conference on Neural Networks*, 2023a.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252. PMLR, 2023b.

Jacopo Talpini, Marco Savi, and Giovanni Neglia. FedBEns: One-shot federated learning based on bayesian ensemble. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=oTCiv1bkjG.

Zhenheng Tang, Yonggang Zhang, Peijie Dong, Yiu-ming Cheung, Amelie C 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.

Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. An aggregation-free federated learning for tackling data heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26233–26242, 2024.

Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. Feddm: Iterative distribution matching for communication-efficient federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16323–16332, 2023.

Guochen Yan, Luyuan Xie, Xinyi Gao, Wentao Zhang, Qingni Shen, Yuejian Fang, and Zhonghai Wu. Fedvck: Non-iid robust and communication-efficient federated learning via valuable condensed knowledge for medical image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21904–21912, 2025.

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*, pp. 6666–6675, 2024.

Dezhong Yao, Yuexin Shi, Tongtong Liu, and Zhiqiang Xu. Fedmho: Heterogeneous one-shot federated learning towards resource-constrained edge devices. *Preprint arXiv:2502.08518*, 2025.

Hu Ye, Jun Zhang, Sibo 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.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. Fedllm-bench: Realistic benchmarks for federated learning of large language models. *Proceedings Advances in Neural Information Processing Systems*, 37:111106–111130, 2024.

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.

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.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6514–6523, 2023.

Yanlin Zhou, George Pu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020.

A Convergence Analysis

Convergence of server-side training on pooled synthetic data

Let the server train on the synthetic dataset $\widetilde{\mathcal{D}} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{m}$ by minimizing

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi) = \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{i=1}^{m} \ell_{\mathrm{CE}}(f_\phi(\tilde{x}_i), \tilde{y}_i) \tag{10}$$

Assumptions

- $\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi)$ is μ-strongly convex and L-smooth.
- The step size of gradient descent is bounded $0 < \eta \leq 1/L$.
- The synthetic distribution $\widetilde{\mathcal{D}}$ is close to the real distribution $\mathcal{D}$ in the sense that for all ϕ,

$$|\mathcal{F}_{\mathcal{D}}(\phi) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi)| \leq \delta. \tag{11}$$

Then, we prove the optimization convergence on the synthetic objective and transfer the bound to the real objective.

Theorem 1: Let $\phi^\star_{\widetilde{\mathcal{D}}} = \arg\min_\phi \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi)$. Gradient descent

$$\phi_{t+1} = \phi_t - \eta \nabla \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) \tag{12}$$

satisfies

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}}) \leq (1 - \eta\mu)^t \big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_0) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big). \tag{13}$$

Moreover, the real data risk is bounded by

$$\mathcal{F}_{\mathcal{D}}(\phi_t) - \mathcal{F}_{\mathcal{D}}(\phi^\star_{\mathcal{D}}) \leq (1 - \eta\mu)^t C_0 + 2\delta, \tag{14}$$

for some constant $C_0 > 0$.

Proof: With the gradient descent:

$$\phi_{t+1} - \phi_t = -\eta \nabla \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t), \tag{15}$$

we substitute into the smoothness inequality result from the descent lemma:

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_{t+1}) \leq \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) + \nabla \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t)^T (\phi_{t+1} - \phi_t) + \frac{L}{2} \|\phi_{t+1} - \phi_t\|^2. \tag{16}$$

to derive

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_{t+1}) \leq \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \eta\big(1 - \frac{\eta L}{2}\big) \|\nabla \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t)\|^2. \tag{17}$$

With the strong convexity

$$\|\nabla \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t)\|^2 \geq 2\mu\big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big), \tag{18}$$

we obtain

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_{t+1}) \leq \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - 2\eta\mu\big(1 - \frac{\eta L}{2}\big)\big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big), \tag{19}$$

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_{t+1}) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}}) \leq \Big(1 - 2\eta\mu\big(1 - \frac{\eta L}{2}\big)\Big)\big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big), \tag{20}$$

with the $\eta \leq 1/L$,

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_{t+1}) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}}) \leq \big(1 - \eta\mu\big)\big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big). \tag{21}$$

By recursion:

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}}) \leq \big(1 - \eta\mu\big)^t\big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_0) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\big). \tag{22}$$

The equation indicates the linear convergence, each step shrinks the error by a factor of $(1 - \eta\mu)$. Following the final assumption at two points:

$$\mathcal{F}_{\mathcal{D}}(\phi_t) \leq \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) + \delta. \tag{23}$$

$$\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}}) \leq \mathcal{F}_{\mathcal{D}}(\phi^\star_{\mathcal{D}}) + \delta. \tag{24}$$

Therefore,

$$\mathcal{F}_{\mathcal{D}}(\phi_t) - \mathcal{F}_{\mathcal{D}}(\phi^\star_{\mathcal{D}}) \leq \Big(\mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_t) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})\Big) + 2\delta. \tag{25}$$

Substituting the gradient descent rate gives

$$\mathcal{F}_{\mathcal{D}}(\phi_t) - \mathcal{F}_{\mathcal{D}}(\phi^\star_{\mathcal{D}}) \leq \big(1 - \eta\mu\big)^t C_0 + 2\delta. \tag{26}$$

where $C_0 = \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi_0) - \mathcal{F}_{\widetilde{\mathcal{D}}}(\phi^\star_{\widetilde{\mathcal{D}}})$.

This establishes that the optimization converges linearly on the synthetic objective, and to a neighborhood of the real-data optimum whose radius is governed by the synthetic-data discrepancy δ.

B Experimental Details

B.1 Datasets

We evaluate our framework on four benchmark datasets: CIFAR-10, CIFAR-100, and SVHN. CIFAR-10 and CIFAR-100 consist of 60,000 RGB images (50,000 train / 10,000 test) at 32×32 resolution across 10 and 100 classes, respectively. SVHN is a real-world digit dataset of 32×32 RGB images partitioned into standard train/test splits.

B.2 Synthetic Data Generation

Each client runs a frozen Stable Diffusion v1.5 pipeline equipped with LCM-LoRA (latent-consistency/lcm-lora-sdv1-5) for accelerated denoising. A CLIP-based IP-Adapter (h94/IP-Adapter) injects visual reference features into the UNet cross-attention layers. The conditioning scale $\gamma \in [0, 1]$ balances adherence to the reference image against text-guided diversity.

B.3 Deviation Evaluation

We evaluate the deviation of synthetic images from their originals using four metrics:

- FID (Fréchet Inception Distance): Measures distributional similarity between synthetic and real images via Inception-v3 feature statistics. Lower FID indicates higher similarity and smaller deviation from the original data.
- Re-ID Rate: Nearest-neighbor retrieval in Inception-v3 feature space measures how often a synthetic image is matched to its original. Higher rate = smaller deviation from the source.

- BLEU Score: Compares captions generated by BLIP for synthetic and nearest original images using n-gram precision n = {1,2,3,4} with brevity penalty. Higher BLEU indicates greater semantic similarity to the original.
- ROUGE-L F1: Measures caption overlap via longest common subsequence (LCS). Higher ROUGE-L F1 implies stronger semantic similarity to the original.

B.4 Implementation Details

All experiments are implemented in PyTorch and conducted on an NVIDIA RTX A6000 (CUDA). To ensure reproducibility, all stochastic processes (data partitioning, model initialization, and generation) are seeded (default seed = 42), and we additionally report averages over seeds {42, 43, 44}. CIFAR and SVHN images are normalized with mean/std (0.5, 0.5, 0.5), while MNIST uses mean 0.1307 and std 0.3081.

B.5 Server Training

The server trains a downstream classifier on the pooled synthetic dataset using SGD with learning rate $\eta = 0.01$, momentum 0.9, and weight decay 5×10^{-4} for 100 epochs (Convnet) with a cosine annealing scheduler and batch size 64. The performance of ResNet18 requires only a single epoch.

C Additional Experiment on Semantic Deviation

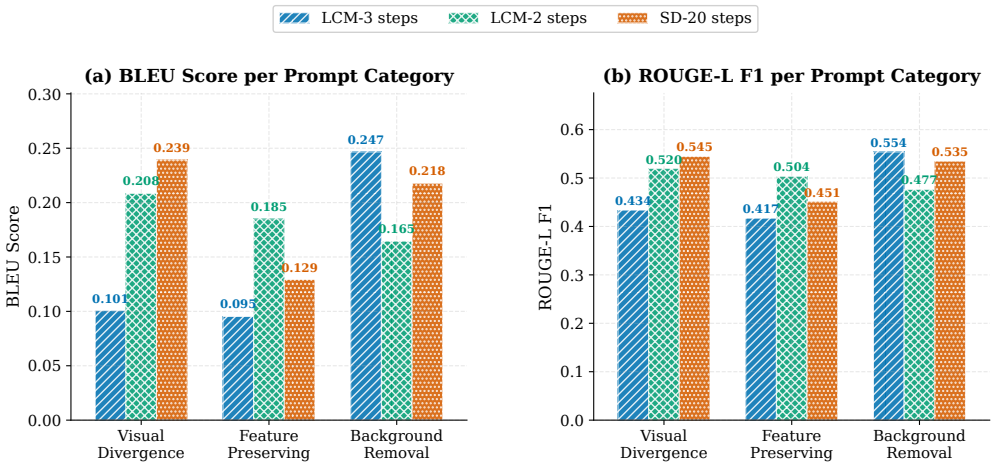

Figure 7: The evaluation of semantic deviation of LCM-LoRA versus Stable Diffusion with two metrics based on BLIP.

To enable a more objective and systematic comparison, we additionally employ BLIP-2 Li et al. (2023) to generate captions for both synthetic and original images, and evaluate their semantic similarity using BLEU Papineni et al. (2002) and ROUGE-L Lin (2004). Because Stable Diffusion with 20 inference steps is computationally expensive, we perform this evaluation on a 0.1 fraction of the dataset for all schemes. As shown in Figure 7, LCM-3-step achieves the lowest BLEU and ROUGE-L scores in the Visual Divergence and Feature Preserving settings, indicating weaker semantic alignment with the original data, consistent with these prompts' goal of departing from the source. In contrast, SD-20 and LCM-2-step show comparable performance. For Background Removal, however, the objective is to preserve the main foreground semantics while modifying only the background; therefore, higher BLEU and ROUGE-L scores indicate better content preservation, and our method achieves the best results in this setting.