LoRA-based Data Augmentation for Federated Learning on Non-IID Data

Anonymous authors
Paper under double-blind review

Abstract

Federated learning (FL) enables privacy-preserving collaborative training. However, its performance degrades substantially under non-independent and identically distributed (non-IID) client data due to client drift or distribution skew. Existing methods often fail to provide distribution-aligned samples to local clients. We propose Federated Harmonization via Diffusion with Low-Rank Adaptation, Federated Harmonization via Diffusion with LoRA, or simply FedHDL, a generative framework that leverages a pretrained Stable Diffusion backbone with lightweight local LoRA adapters. Instead of aggregating full parameters, the server aggregates only LoRA adapters and performs cross-style synthesis to generate missing class–style samples. By using these synthetic samples for data augmentation, FedHDL effectively aligns local data with the global distribution, thereby re-balances label distributions and mitigates feature skew while maintaining low communication overhead. Extensive experiments across diverse non-IID scenarios show that FedHDL consistently improves upon prior generative augmentation methods while achieving competitive performance against strong federated optimization and personalization baselines. The experimental results demonstrate that parameter-efficient generative alignment is an effective path toward robust federated learning under distribution shift.

1 Introduction

Federated Learning (FL) McMahan et al. (2017) enables multiple clients to collaboratively train a shared model without centralizing private data. The mechanisms make FL attractive for privacy-sensitive applications. The applications may include those working on smart agentic devices around us, personal records like medical data, and many others where data-collecting organizations do not attempt or not allowed to share the on-site data with others. However, FL often suffers from substantial accuracy degradation and unstable convergence when client data is non-IID Kairouz et al. (2021); Luo et al. (2021); Shi et al. (2023). In practice, each client observes only a narrow slice of the global distribution that could vary in classes, backgrounds, or visual domains among different clients, which leads to severe *label skew* and *feature skew* in those local data. Such heterogeneity induces client drift, where local updates become inconsistent, ultimately undermining global generalization after aggregation Zhao et al. (2018); Zhu et al. (2021).

To mitigate these issues, a large body of work has focused on algorithmic refinements, including specialized aggregation schemes, objective regularization Li et al. (2020); Karimireddy et al. (2020), and strategic client selection. Although these methods offer incremental improvements by stabilizing the optimization trajectory, they often overlook the fundamental root of the non-IID problem: the inherent inconsistency between local and global data distributions. These optimization-based remedies do not resolve a critical limitation—clients lack training evidence representative of the global distribution. Without more distribution-aligned local data, adjusting the optimization path alone can only partially mitigate the impact of severe distribution shifts.

Data augmentation has emerged as a compelling alternative to address this data-level deficiency. However, conventional techniques often fail to generate missing class–style combinations that are completely absent from individual clients Yoon et al. (2021); Azizi et al. (2023). Although some recent work attempts to bridge these gaps using auxiliary public datasets or shared synthetic data Bansal & Grover (2023), such public data are frequently unavailable or introduce significant

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.

domain mismatch, while sharing raw synthesized samples raises persistent privacy concerns and substantial communication overhead.

Motivated by these gaps, we propose to resolve the non-IID problem at its source by aligning local data distributions with the global distribution via *generative harmonization*. We introduce FedHDL (Federated Harmonization via Diffusion with Low-Rank Adaptation or Federated Harmonization via Diffusion with LoRA), a parameter-efficient framework that leverages pretrained Stable Diffusion (SD) Rombach et al. (2022) to perform style-aware data augmentation. FedHDL attaches lightweight *LoRA adapters* to a shared SD backbone, allowing clients to locally capture unique style signatures with minimal communication cost. By performing *cross-style synthesis* at the server level, FedHDL provides each client with distribution-aligned training samples that bridge domain gaps and enhance sample diversity. We evaluate FedHDL on four benchmarks—CIFAR-10, Office-Home, PACS, and DomainNet—covering both label and feature heterogeneity.

We summarize the contributions of this work as follows:

1.1 Our Contributions

- We identify a fundamental limitation in optimization-based FL: the inability to provide distribution-aligned training samples for missing class–style combinations. To address this, we propose FedHDL, a generative framework that leverages the Stable Diffusion and LoRA computation to synthesize client-specific missing samples.
- FedHDL provides a unified solution for focused data that are with diverse heterogeneity, effectively mitigating both label skew and feature skew simultaneously. The proposed parameter-efficient strategy bridges domain gaps at the data level, aligning local distributions with the global goal while significantly enhancing sample diversity.
- Extensive experiments on four benchmarks show that FedHDL consistently improves over prior generation-based augmentation methods and achieves competitive performance against strong optimization- and personalization-based FL baselines across diverse non-IID settings.

Next, we step-by-step present the proposed method in the following sections. Right after the introduction, we discuss the prior work that is related to this research in Section 2. Following that, we demonstrate the proposed method in detail in Section 3, which is followed by its evaluation by a set of experiments in Section 4. In the end, we conclude the work in Section 5.

2 Related Work

In this section, we survey the prior work that is related to several topics. First, we discuss how non-IID data can introduce severe performance degrading in FL, and the past efforts on mitigating this situation. Following that, we review some past data augmentation algorithms and diffusion models that can be used in FL scenarios.

2.1 Non-IID in Federated Learning

Non-IID data distributions are a primary obstacle in federated learning (FL), often leading to client drift and degraded convergence under both label skew and feature/domain skew McMahan et al. (2017); Kairouz et al. (2021). FedAvg McMahan et al. (2017) serves as the canonical baseline that aggregates client models by weighted averaging. However, it is prone to drift when local objectives are highly inconsistent. A large body of work mitigates this issue from the optimization perspective by stabilizing local updates or reducing the variance induced by heterogeneous clients. FedProx Li et al. (2020) introduces a proximal term during local training to constrain the deviation of local model weights from the global model, improving stability when client objectives differ substantially. SCAFFOLD Karimireddy et al. (2020) corrects client drift using control variates that reduce the mismatch between local and global gradients, thereby accelerating convergence under heterogeneity. FedNova Wang et al. (2020) and FedSOL Lee et al. (2024) normalize client updates to account for varying numbers of local steps, mitigating objective inconsistency caused by heterogeneous local computation. FedSAM Caldarola et al. (2022) incorporates Sharpness-Aware Minimization (SAM)

into federated training to favor flatter minima, enhancing generalization and stability under non-IID settings.

Beyond optimization, several approaches explicitly target label heterogeneity. FedNTD Lee et al. (2022) uses non-target distillation to align clients on the relative probabilities of incorrect classes, reducing prediction bias under label skew. For personalization, FedPer Arivazhagan et al. (2019) shares lower layers while keeping client-specific classification layers, which is effective when clients exhibit different label distributions. To address feature/domain skew, FedBN Li et al. (2021) keeps batch-normalization statistics local, which helps when clients have different visual styles or acquisition conditions. DualFed Zhu et al. (2024) further improves robustness by decoupling and jointly optimizing global and personalized components, covering both label and feature shifts. Other lines tackle feature skew by enforcing domain-invariant representations or feature alignment across clients; however, these methods still rely on the existing local data distribution and do not explicitly provide distribution-aligned samples to fill missing class–style combinations for each client.

2.2 Data Augmentation in Federated Learning

Creating IID-like training data via data augmentation has emerged as a compelling alternative to optimization-only strategies. Traditional approaches either leverage auxiliary public datasets to rebalance client training Zhao et al. (2018) or depend on communicating condensed/synthetic data across the network Hu et al. (2022b); Shipard et al. (2023); Goetz & Tewari (2020); Morafah et al. (2024); Jeong et al. (2018). However, public data are often unavailable or introduce significant domain mismatch, while directly sharing samples raises substantial privacy and communication concerns. FedHDL departs from these methods by performing high-quality generative augmentation with a pretrained Stable Diffusion backbone, exchanging only lightweight LoRA adapters without directly sharing raw client data.

2.3 Diffusion Models and Parameter-Efficient Adaptation

Diffusion models Rombach et al. (2022); Cao et al. (2024) have redefined the state-of-the-art in generative modeling, offering superior stability and sample quality compared to GANs Ho et al. (2020); Song et al. (2020); Dhariwal & Nichol (2021). Large-scale systems like Stable Diffusion enable controllable image synthesis via language prompts. Recently, Low-Rank Adaptation (LoRA) Hu et al. (2022a) has become the de facto standard for parameter-efficient fine-tuning (PEFT), as it allows massive backbones to be adapted with minimal overhead.

2.4 Generative Harmonization and the S2R Gap

A primary challenge in *synthetic data generation* is the *synthesis-to-real (S2R) gap*, where discrepancies in texture and style limit the utility of synthetic data for downstream tasks Azizi et al. (2023); He et al. (2022). Existing diffusion-based FL methods often treat generation as purely class-conditional and overlook *domain/style skew*. FedHDL addresses this by coupling federated LoRA adaptation with *cross-style synthesis*. By utilizing domain-specific adapters to tailor synthetic images to a client's target distribution, our framework effectively narrows the S2R gap and improves both global and local performance under real-world heterogeneity, such as the Office-Home benchmark.

3 Methodology

3.1 Stage 1: Federated LoRA Adaptation

To capture client-specific style distributions while preserving privacy, we leverage Low-Rank Adaptation (LoRA) Hu et al. (2022a) within the federated environment. Each client k has private datasets $\mathcal{D}_k$ characterized by unique domain or label skews.

Local LoRA Training Each client k initializes a pretrained Stable Diffusion U-Net θ_0 and inserts LoRA adapters $\phi_k = \{(\mathbf{A}_k^\ell, \mathbf{B}_k^\ell)\}_{\ell \in \mathcal{L}}$ into key attention modules. During local training, θ_0 remains

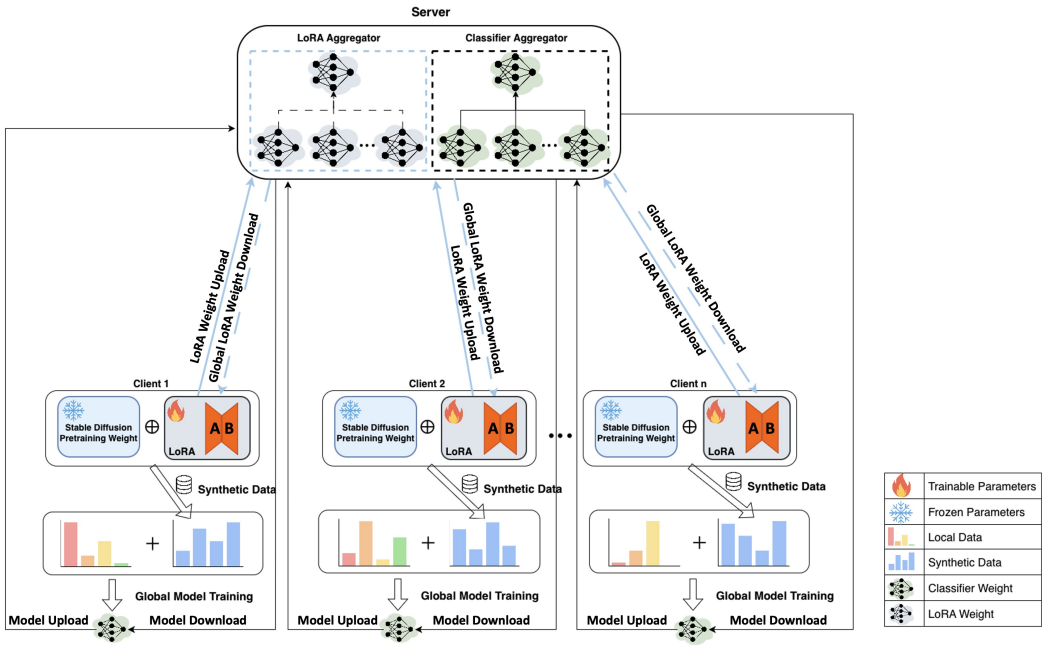


Figure 1: Overview of the **FedHDL** framework. Step 1: Clients fine-tune local LoRA adapters on private data. Step 2: The server performs style-conditioned cross-style synthesis to balance labels and align distributions. Step 3: Clients train the final federated classifier using harmonized datasets.

frozen while ϕ_k is optimized for E epochs on private data $\mathcal{D}_k$:

$$\phi_k \leftarrow \arg\min_{\phi} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_k} \left[\mathcal{L}_{\text{DM}}(\theta_0;\phi)\right]. \tag{1}$$

Only the LoRA adapter parameters are transmitted to the server, avoiding direct sharing of raw client data while reducing communication overhead.

Weight-Space SVD Aggregation Directly averaging LoRA factors can lead to unstable updates due to *factorization invariance*; specifically, for a global update,

$$\sum_k \frac{n_k}{n}(\mathbf{B}_k\mathbf{A}_k) \neq \left(\sum_k \frac{n_k}{n}\mathbf{B}_k\right)\left(\sum_k \frac{n_k}{n}\mathbf{A}_k\right). \tag{2}$$

This mathematical misalignment motivates us to aggregate in the induced weight-update space: $\Delta\mathbf{W}_k^\ell = \mathbf{B}_k^\ell\mathbf{A}_k^\ell$. The server computes a weighted average $\overline{\Delta\mathbf{W}}^\ell = \sum_k \frac{n_k}{n}\Delta\mathbf{W}_k^\ell$ and factorizes it into rank-r components via truncated SVD:

$$\overline{\Delta\mathbf{W}}^\ell \approx \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top \implies \mathbf{B}_{\text{global}}^\ell = \mathbf{U}_r\mathbf{\Sigma}_r^{1/2},\ \mathbf{A}_{\text{global}}^\ell = \mathbf{\Sigma}_r^{1/2}\mathbf{V}_r^\top. \tag{3}$$

3.2 Stage 2: Cross-Style Synthesis and Harmonization

To mitigate feature heterogeneity, FedHDL aligns the class-conditional feature distributions $p(\mathbf{x} \mid y)$ across clients through cross-style synthesis. The server maintains a global pool of style-specific LoRA adapters, denoted by $\{\phi^{(s)}\}_{s=1}^{S}$, where each adapter captures the visual characteristics of a particular client domain or style. Instead of restricting generation to the style observed locally, each client can use adapters associated with other domains to synthesize class-preserving samples with previously unseen visual characteristics. This expands the support of the local feature distribution and encourages the classifier to learn representations that are less dependent on client-specific styles.

To reduce label imbalance simultaneously, we adopt a max-label augmentation strategy. For client k, let

$$N_{k,\max} = \max_y |\mathcal{D}_{k,y}|, \tag{4}$$

where $\mathcal{D}_{k,y}$ denotes the subset of local samples belonging to class y. For each underrepresented class satisfying $|\mathcal{D}_{k,y}| < N_{k,\max}$, the required number of synthetic samples is defined as

$$N_{k,y}^{\text{syn}} = N_{k,\max} - |\mathcal{D}_{k,y}|. \tag{5}$$

The client then generates a synthetic subset $\mathcal{D}_{k,y}^{\text{syn}}$ containing $N_{k,y}^{\text{syn}}$ samples, such that the resulting class counts become balanced within the client. The complete synthetic dataset and augmented local dataset are given by

$$\mathcal{D}_k^{\text{syn}} = \bigcup_y \mathcal{D}_{k,y}^{\text{syn}}, \qquad \widetilde{\mathcal{D}}_k = \mathcal{D}_k \cup \mathcal{D}_k^{\text{syn}}. \tag{6}$$

This strategy supplements missing or underrepresented labels while introducing diverse domain styles, thereby addressing both label skew and feature skew at the data level.

Personalized Generation Directly applying a style adapter from another client may produce samples that are visually diverse but insufficiently aligned with the local task distribution. To balance local relevance and cross-domain diversity while accounting for the non-uniqueness of LoRA factorization, FedHDL performs interpolation in the induced weight-update space rather than directly interpolating the LoRA factors.

For each LoRA-injected layer ℓ, the updates induced by the global and selected style-specific adapters are respectively defined as

$$\Delta W_{\text{global}}^{\ell} = B_{\text{global}}^{\ell} A_{\text{global}}^{\ell}, \qquad \Delta W^{(s),\ell} = B^{(s),\ell} A^{(s),\ell}. \tag{7}$$

FedHDL constructs the personalized weight update as

$$\widetilde{\Delta W}_{k,s}^{\ell} = (1-\lambda_k)\Delta W_{\text{global}}^{\ell} + \lambda_k \Delta W^{(s),\ell}, \qquad \lambda_k \in [0,1]. \tag{8}$$

Because the interpolated update may have a rank of up to $2r$, truncated SVD is applied to obtain its best rank-r approximation:

$$\widetilde{\Delta W}_{k,s}^{\ell} \approx U_r^{\ell} \Sigma_r^{\ell} (V_r^{\ell})^{\top}. \tag{9}$$

The personalized LoRA factors are then reconstructed as

$$B_{k,s}^{\ell} = U_r^{\ell} (\Sigma_r^{\ell})^{1/2}, \qquad A_{k,s}^{\ell} = (\Sigma_r^{\ell})^{1/2} (V_r^{\ell})^{\top}. \tag{10}$$

The resulting personalized adapter is defined as $\phi_{k,s} = \{A_{k,s}^{\ell}, B_{k,s}^{\ell}\}_{\ell}$. By construction, $B_{k,s}^{\ell} A_{k,s}^{\ell} \approx \widetilde{\Delta W}_{k,s}^{\ell}$, ensuring that the interpolation is invariant to equivalent LoRA factorizations and consistent with the weight-space aggregation described above. Here, ϕ_{global} captures shared generative knowledge across clients, $\phi^{(s)}$ represents the target style, and λ_k controls the strength of style adaptation. A smaller λ_k preserves more globally shared characteristics, whereas a larger value emphasizes the selected target style.

For each class y, the reconstructed personalized adapter $\phi_{k,s}$ is combined with a class- and style-conditioned prompt to generate samples that preserve the target class semantics while incorporating styles absent from $\mathcal{D}_k$. Through this process, the augmented distribution $\widetilde{\mathcal{D}}_k$ covers a broader range of visual appearances and more closely approximates the global conditional distribution $p_{\text{global}}(\mathbf{x} \mid y)$. Consequently, local classifiers are trained on more balanced and stylistically diverse data, resulting in model updates that are more robust to domain shifts and more consistent with the global learning objective.

3.3 Stage 3: Federated Classifier Training

In the final stage, each client trains a local classifier using its augmented dataset, which combines the original private samples with the synthetic samples generated during the harmonization stage. For client k, the augmented training set is defined as

$$\widetilde{\mathcal{D}}_k = \mathcal{D}_k \cup \mathcal{D}_k^{\text{syn}}, \tag{11}$$

where $\mathcal{D}_k$ denotes the original local dataset and $\mathcal{D}_k^{\text{syn}}$ represents the distribution-aligned synthetic dataset. Starting from the global classifier parameters received from the server, each participating

client performs local stochastic gradient descent on $\widetilde{\mathcal{D}}_k$ to obtain the updated local classifier parameters θ_k. The local optimization objective can be written as

$$\min_{\theta_k} \frac{1}{|\widetilde{\mathcal{D}}_k|} \sum_{(x,y)\in\widetilde{\mathcal{D}}_k} \ell\big(f_{\theta_k}(x), y\big), \tag{12}$$

where f_{θ_k} denotes the local classifier and $\ell(\cdot)$ is the classification loss. By jointly training on real and synthetic samples, the local classifier is exposed to a broader and more representative label and feature distribution than that provided by the original client dataset alone.

A direct application of conventional FedAvg would assign aggregation weights according to the total number of local training samples. However, treating real and synthetic samples as equally reliable may overestimate the contribution of generated data. Although the synthetic samples are designed to compensate for missing classes or domain styles, they may still contain visual artifacts, imperfect semantic details, or a mild synthetic-to-real distribution gap. Consequently, a client containing a large number of generated samples could receive an excessively large aggregation weight even when the additional samples provide less reliable supervision than real observations.

To address this issue, we introduce *Synthetic-Confidence-Weighted FedAvg* (SynCW-FedAvg), which explicitly distinguishes between the contributions of real and synthetic samples. For each client, we define an effective dataset size as

$$n_k^{\text{eff}} \triangleq |\mathcal{D}_k| + \alpha|\mathcal{D}_k^{\text{syn}}|, \qquad \alpha \in [0,1], \tag{13}$$

where α denotes the confidence assigned to each synthetic sample. The original real samples are implicitly assigned a weight of 1, whereas each generated sample contributes only a fraction α to the effective dataset size. Therefore, the aggregation weight of a client depends not only on the quantity of its augmented data but also on the estimated reliability of the synthetic portion.

Given the set of participating clients $\mathcal{S}$ in the current communication round, the server updates the global classifier according to

$$\theta \leftarrow \sum_{k\in\mathcal{S}} \frac{n_k^{\text{eff}}}{\sum_{j\in\mathcal{S}} n_j^{\text{eff}}} \theta_k. \tag{14}$$

Equivalently, the normalized aggregation coefficient for client k is given by

$$w_k = \frac{|\mathcal{D}_k| + \alpha|\mathcal{D}_k^{\text{syn}}|}{\sum_{j\in\mathcal{S}} \big(|\mathcal{D}_j| + \alpha|\mathcal{D}_j^{\text{syn}}|\big)}, \tag{15}$$

such that $\sum_{k\in\mathcal{S}} w_k = 1$. SynCW-FedAvg discounts synthetic samples through the confidence parameter α, preventing clients with large synthetic datasets from disproportionately dominating global aggregation. Here, $\alpha = 0$ ignores synthetic samples in aggregation weighting, whereas $\alpha = 1$ treats real and synthetic samples equally. Empirically, $\alpha = 0.6$ yields the best performance and is used by default. This reliability-aware aggregation allows FedHDL to exploit synthetic augmentation while reducing the impact of imperfect generated samples under non-IID settings.

4 Experiments

Experimental Setup. We evaluate FedHDL across multiple non-IID settings and compare it with representative federated learning baselines. For label heterogeneity, we use CIFAR-10 with Dirichlet partitions $\beta \in \{0.05, 0.1, 0.3, 0.5\}$. For feature heterogeneity, we evaluate on Office-Home, PACS, and DomainNet using domain-partitioned clients. Depending on the setting, we compare against optimization-based, personalization-based, domain-adaptive, and generation-based FL methods, while FedAvg and FedProx serve as common reference baselines across scenarios. All results are averaged over three runs. Full details of dataset partitions, model architectures, training hyperparameters, baseline configurations, LoRA settings, and diffusion-based image generation are provided in the supplementary material.

4.1 Results on Label Heterogeneity

To evaluate the robustness of FedHDL against label heterogeneity, we conduct extensive comparisons with state-of-the-art federated learning baselines on the CIFAR-10 dataset using a Dirichlet

Table 1: Test accuracy (%) on CIFAR-10 under different Dirichlet label-skew levels. Results are reported as mean ± std over three runs. Best and second-best federated results are shown in **bold** and underline, respectively.

	Dirichlet Label Skew			
Method	$\boldsymbol{\beta = 0.05}$	$\boldsymbol{\beta = 0.1}$	$\boldsymbol{\beta = 0.3}$	$\boldsymbol{\beta = 0.5}$
Central	76.09 ± 0.09	76.58 ± 0.10	76.78 ± 0.34	76.39 ± 0.27
FedAvg McMahan et al. (2017)	42.27 ± 1.34	56.13 ± 0.78	67.32 ± 0.94	73.90 ± 0.66
FedProx Li et al. (2020)	50.58 ± 0.57	59.80 ± 1.12	68.39 ± 0.81	72.87 ± 0.55
FedNova Wang et al. (2020)	10.00 *(Failed)*	10.00 *(Failed)*	64.67 ± 0.77	70.04 ± 0.45
Scaffold Karimireddy et al. (2020)	10.00 *(Failed)*	10.00 *(Failed)*	71.92 ± 0.17	$\underline{75.49 \pm 0.21}$
FedNTD Lee et al. (2022)	58.08 ± 0.48	63.16 ± 1.02	71.56 ± 0.26	74.91 ± 0.33
FedSAM Qu et al. (2022)	36.14 ± 1.21	52.14 ± 0.94	64.83 ± 0.56	70.74 ± 0.40
FedASAM Caldarola et al. (2022)	43.12 ± 1.25	57.00 ± 0.30	67.45 ± 0.92	73.91 ± 0.51
FedSOL Lee et al. (2024)	$\underline{60.01 \pm 0.30}$	$\underline{64.13 \pm 0.46}$	$\underline{71.94 \pm 0.57}$	$\mathbf{75.60 \pm 0.32}$
FedHDL (Ours)	$\mathbf{71.41 \pm 0.76}$	$\mathbf{71.57 \pm 0.56}$	$\mathbf{73.64 \pm 0.55}$	73.36 ± 0.43

partition strategy with $\beta \in \{0.05, 0.1, 0.3, 0.5\}$. As summarized in Table 1, FedHDL achieves strong overall performance across different levels of label skew, with its advantage becoming most pronounced under severe heterogeneity.

Notably, our performance gains become increasingly pronounced as the label skew intensifies. Under extreme non-IID conditions ($\beta = 0.05$), where traditional methods such as FedNova and Scaffold fail to converge effectively (reporting only 10.00% accuracy), FedHDL achieves $71.41 \pm 0.76\%$ accuracy, outperforming the second-best baseline FedSOL by nearly 11.4%. These results demonstrate that while optimization-only strategies struggle with severe client drift, FedHDL's generative augmentation effectively aligns local client distributions with the global distribution by synthesizing missing class–style samples. By providing distribution-aligned samples, These results highlight the benefit of synthesizing missing-class samples to mitigate client drift under severe label heterogeneity.

4.2 Results on Feature Heterogeneity

Tables 2–4 compare FedHDL with representative FL methods under domain-induced feature heterogeneity. FedHDL consistently improves over the generation-based augmentation baseline GenFedSD across all three benchmarks, increasing average accuracy from 77.99% to 80.53% on Office-Home, from 93.32% to 95.00% on PACS, and from 83.72% to 84.35% on DomainNet. These results indicate that federated LoRA adaptation and cross-style harmonization provide more effective distribution-aligned augmentation than directly applying diffusion-based synthetic data.

FedHDL also remains competitive with strong optimization and personalization-based FL methods. It achieves the best average accuracy on Office-Home ($80.53 \pm 0.25\%$), outperforming DualFed by 0.79 points, and nearly matches DualFed on PACS (95.00% vs. 95.01%). On the more challenging DomainNet benchmark, FedHDL reaches $84.35 \pm 0.09\%$, compared with $86.14 \pm 0.12\%$ for DualFed, while obtaining strong results on several individual domains. Overall, FedHDL substantially improves upon prior generative augmentation and achieves performance comparable to strong FL baselines without relying solely on optimization-level corrections.

4.3 Scalability Analysis

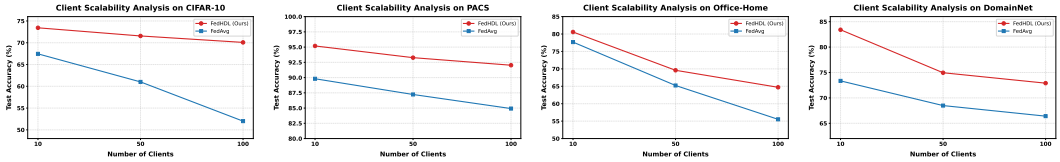
As shown in Fig. 2, FedHDL consistently outperforms FedAvg as the number of clients increases from 10 to 100, with five clients participating per round. Despite the resulting drop in participation rate from 50% to 5%, FedHDL remains more robust across both label- and feature-heterogeneous settings. At 100 clients, FedHDL achieves 70.08%, 92.03%, 64.71%, and 72.90% on CIFAR-10, PACS, Office-Home, and DomainNet, respectively, compared with 52.00%, 84.92%, 55.52%, and 66.41% for FedAvg. These results indicate that FedHDL scales more effectively under increasingly fragmented and sparse client participation.

Table 2: Test accuracy (%) on Office-Home under feature heterogeneity. Results are reported as mean ± std over three runs. Best and second-best federated results are shown in **bold** and underline, respectively.

	Office-Home Domains				
Method	**A**	**C**	**P**	**R**	**Avg.**
SingleSet	66.52 ± 1.27	74.27 ± 0.60	87.46 ± 1.02	77.54 ± 0.58	76.45 ± 0.32
FedAvg McMahan et al. (2017)	68.82 ± 1.30	74.91 ± 1.02	85.82 ± 0.36	80.30 ± 0.53	77.46 ± 0.35
FedProx Li et al. (2020)	68.78 ± 1.37	74.73 ± 0.79	85.73 ± 0.35	80.25 ± 0.70	77.37 ± 0.33
FedPer Arivazhagan et al. (2019)	70.31 ± 1.07	75.03 ± 0.38	87.76 ± 0.18	80.51 ± 0.43	78.40 ± 0.40
FedRep Collins et al. (2021)	70.23 ± 0.96	75.44 ± 0.69	85.82 ± 0.45	80.39 ± 0.92	77.97 ± 0.37
FedBN Li et al. (2021)	68.58 ± 1.23	76.01 ± 0.45	86.31 ± 0.96	79.40 ± 0.40	77.58 ± 0.29
FedETF Li et al. (2023)	69.90 ± 1.14	74.64 ± 0.41	85.52 ± 0.35	80.18 ± 0.39	77.56 ± 0.29
DualFed Zhu et al. (2024)	71.01 ± 0.71	$\mathbf{77.41 \pm 0.47}$	88.84 ± 0.47	81.70 ± 0.28	79.74 ± 0.37
Gen-FedSD Morafah et al. (2024)	68.11 ± 0.09	72.34 ± 0.09	89.25 ± 0.10	82.26 ± 0.11	77.99 ± 0.06
FedHDL (Ours)	$\mathbf{74.63 \pm 1.31}$	73.36 ± 0.43	$\mathbf{91.71 \pm 0.34}$	$\mathbf{82.41 \pm 0.23}$	$\mathbf{80.53 \pm 0.25}$

Table 3: Test accuracy (%) on PACS under feature heterogeneity. Results are reported as mean ± std. For Gen-FedSD, statistics are computed over the final three communication rounds. Best and second-best federated results are shown in **bold** and underline, respectively.

	PACS Domains				
Method	**P**	**A**	**C**	**S**	**Avg.**
SingleSet	97.78 ± 0.56	88.12 ± 0.25	89.19 ± 0.37	91.01 ± 0.73	91.52 ± 0.10
FedAvg McMahan et al. (2017)	97.72 ± 0.56	89.24 ± 1.01	89.32 ± 0.60	91.01 ± 0.70	91.82 ± 0.34
FedProx Li et al. (2020)	97.90 ± 0.38	89.14 ± 1.18	89.40 ± 0.61	91.52 ± 0.72	91.99 ± 0.38
FedPer Arivazhagan et al. (2019)	98.20 ± 0.42	89.54 ± 1.16	91.28 ± 0.75	91.29 ± 0.60	92.58 ± 0.57
FedRep Collins et al. (2021)	97.84 ± 0.35	89.83 ± 1.33	89.96 ± 0.27	91.39 ± 0.48	92.25 ± 0.22
FedBN Li et al. (2021)	92.20 ± 0.46	89.88 ± 0.86	90.38 ± 0.75	91.34 ± 0.53	92.45 ± 0.37
FedETF Li et al. (2023)	97.43 ± 0.24	90.95 ± 0.77	90.26 ± 0.29	90.70 ± 0.68	92.33 ± 0.30
DualFed Zhu et al. (2024)	98.32 ± 0.24	92.47 ± 0.42	$\mathbf{94.91 \pm 0.63}$	$\mathbf{94.32 \pm 0.61}$	$\mathbf{95.01 \pm 0.31}$
Gen-FedSD Morafah et al. (2024)	98.92 ± 0.27	93.58 ± 0.46	92.50 ± 0.70	88.30 ± 2.65	93.32 ± 0.98
FedHDL (Ours)	$\mathbf{99.16 \pm 0.13}$	$\mathbf{95.05 \pm 0.55}$	94.20 ± 0.09	91.58 ± 0.28	95.00 ± 0.06



(a) CIFAR-10 (b) PACS (c) Office-Home (d) DomainNet

Figure 2: Client scalability comparison between FedHDL and FedAvg across four benchmarks as the number of clients increases from 10 to 100, with five participating clients per communication round.

4.4 Ablation Study

To evaluate the contribution of each component in FedHDL, we perform an ablation study on CIFAR-10 under extreme label skew. As shown in Table 5, incorporating generative cross-style synthesis yields the most substantial performance leap, elevating average accuracy from $42.27 \pm 1.34\%$ to $67.77 \pm 1.34\%$. This validates that synthesizing distribution-aligned samples fundamentally addresses the missing-class problem. Integrating weight-space SVD aggregation further boosts accuracy to $69.49 \pm 2.29\%$, confirming that low-rank feature extraction effectively filters out local domain noise from heterogeneous adapters. Finally, incorporating the synthetic-confidence weight achieves the optimal performance of $71.41 \pm 0.76\%$. This underscores that dynamically balancing the contributions of real and synthetic data based on effective sample sizes provides an additional, complementary benefit to model generalization.

Table 4: Test accuracy (%) on DomainNet under feature heterogeneity. Results are reported as mean ± std. For Gen-FedSD, statistics are computed over the final three communication rounds. Best and second-best federated results are shown in **bold** and underline, respectively.

	DomainNet Domains						
Method	**Clipart**	**Infograph**	**Painting**	**Quickdraw**	**Real**	**Sketch**	**Avg.**
SingleSet	88.25 ± 0.81	50.99 ± 1.24	89.60 ± 1.00	82.78 ± 0.43	94.07 ± 0.12	88.16 ± 0.53	82.31 ± 0.19
FedAvg McMahan et al. (2017)	89.47 ± 0.97	53.70 ± 0.86	89.60 ± 0.52	80.58 ± 0.80	92.85 ± 0.52	88.56 ± 0.58	82.46 ± 0.33
FedProx Li et al. (2020)	89.47 ± 0.86	53.79 ± 0.96	89.56 ± 0.55	80.56 ± 0.87	92.87 ± 0.54	88.63 ± 0.56	82.48 ± 0.38
FedPer Arivazhagan et al. (2019)	89.70 ± 0.81	54.22 ± 0.68	$\underline{92.12 \pm 0.98}$	82.18 ± 0.65	94.76 ± 0.41	89.57 ± 0.66	83.76 ± 0.32
FedRep Collins et al. (2021)	89.62 ± 0.76	54.19 ± 0.71	90.60 ± 0.37	80.84 ± 0.91	93.03 ± 0.49	89.03 ± 0.77	82.88 ± 0.24
FedBN Li et al. (2021)	89.85 ± 0.67	54.58 ± 1.04	91.34 ± 0.90	80.62 ± 0.68	93.76 ± 0.44	89.06 ± 0.41	83.20 ± 0.36
FedETF Li et al. (2023)	88.97 ± 0.81	55.65 ± 0.85	91.76 ± 0.52	79.76 ± 0.48	94.15 ± 0.23	89.03 ± 0.39	83.22 ± 0.31
DualFed Zhu et al. (2024)	$\mathbf{92.51 \pm 0.41}$	56.77 ± 0.95	$\mathbf{94.41 \pm 0.30}$	$\mathbf{85.18 \pm 0.30}$	94.69 ± 0.08	$\mathbf{92.27 \pm 0.54}$	$\mathbf{86.14 \pm 0.12}$
Gen-FedSD Morafah et al. (2024)	89.93 ± 0.37	$\mathbf{62.76 \pm 0.25}$	90.56 ± 0.11	72.33 ± 0.07	$\underline{95.83 \pm 0.03}$	90.91 ± 0.07	83.72 ± 0.08
FedHDL (Ours)	$\underline{91.18 \pm 0.27}$	$\underline{61.57 \pm 0.85}$	91.16 ± 0.08	74.77 ± 1.27	$\mathbf{95.93 \pm 0.04}$	$\underline{91.50 \pm 0.78}$	$\underline{84.35 \pm 0.09}$

Table 5: Ablation study of FedHDL components on CIFAR-10. Results are reported as best accuracy (mean ± std).

Synthesis	**SVD**	**Synthetic-Confidence Weight**	**Accuracy (%)**
✗	✗	✗	42.27 ± 1.34
✓	✗	✗	67.77 ± 1.34
✓	✓	✗	69.49 ± 2.29
✓	✓	✓	$\mathbf{71.41 \pm 0.76}$

4.5 Limitations

When the label distribution approaches IID ($\beta = 0.5$), the benefit of data harmonization decreases because clients already contain relatively diverse class samples. Additional synthetic data may introduce redundancy or mild distribution mismatch while incurring extra computation and communication costs. Future work will explore adaptive augmentation policies that dynamically adjust the synthesis quota based on the degree of client heterogeneity.

5 Conclusions

In this work, we proposed FedHDL, a generative data harmonization framework for mitigating data heterogeneity in federated learning. By combining a pretrained Stable Diffusion model with parameter-efficient LoRA adapters, FedHDL augments local clients with distribution-aligned synthetic samples without requiring auxiliary public data. Extensive experiments on four benchmarks—CIFAR-10, Office-Home, PACS, and DomainNet—covering both label and feature heterogeneity demonstrate the general applicability of our approach. FedHDL consistently improves upon prior generation-based data augmentation methods and achieves performance competitive with strong optimization- and personalization-based FL methods across diverse non-IID settings. These results suggest that addressing heterogeneity at the data-distribution level provides a complementary direction to conventional optimization-based approaches. Future work will explore more adaptive generation strategies and efficient adapter architectures to further improve generalization and scalability.

Reproducibility Statement

To support reproducibility, we provide detailed descriptions of the datasets, data partitions, model architectures, baseline configurations, optimization hyperparameters, LoRA settings, and diffusion-based image generation procedures in the supplementary material. We also report the experimental protocols, random seeds, and evaluation settings used throughout our experiments.

AI USE STATEMENT

Generative AI tools were used to assist with code development and debugging, and language editing. All AI-assisted code, analyses, and written content were manually reviewed and verified by the authors, and all experimental results reported in this work were obtained from our own implementations and experiments. The authors take full responsibility for the final content, claims, and results presented in this paper.

REFERENCES

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023.

Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision (ECCV)*, pp. 654–672, Cham, 2022. Springer Nature Switzerland.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Liam Collins et al. Exploiting shared representations for personalized federated learning. *PMLR, 2021. p. 2089-2099*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Jack Goetz and Ambuj Tewari. Federated learning via synthetic data. *arXiv preprint arXiv:2008.04489*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Tzu-Ming Harry Hsu et al. Measuring the effects of non-identical data distribution for federated visual classification. In *arXiv preprint arXiv:1909.06335*, 2019.

Edward J. Hu et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022a.

Shengyuan Hu et al. Fedsynth: Gradient compression via synthetic data in federated learning. *arXiv preprint arXiv:2204.01273*, 2022b.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy et al. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 2020.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *University of Toronto*, 7, 2009.

Gihun Lee, Minchan Jeong, Young-Duk Shin, Sang-Hoon Bae, and Se-Young Yun. Preservation of the global knowledge by not-true distillation in federated learning. In *NeurIPS*, volume 35, pp. 38461–38474, 2022.

Gihun Lee, Minchan Jeong, Sang-Ho Kim, Jong-Hoon Oh, and Se-Young Yun. Fedsol: Stabilized orthogonal learning with proximal restrictions in federated learning. In *CVPR*, pp. 12512–12522, 2024.

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*, pp. 5542–5550, 2017.

Tian Li et al. Federated optimization in heterogeneous networks. In *Proceedings of Machine learning and systems*, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Zexi Li et al. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. *In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5319-5329)*, 2023.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 5972–5984, 2021.

Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics. PMLR*, 2017.

Mahdi Morafah, Matthias Reisser, Bill Lin, and Christos Louizos. Stable diffusion-based data augmentation for federated learning with non-iid data. *arXiv preprint arXiv:2405.07925*, 2024.

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*, pp. 1406–1415, 2019.

Zhe Qu, Xingyu Li, Rui Duan, Gongsheng Nie, and Cheng Luo. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning (ICML)*, pp. 18250–18280. PMLR, 2022.

Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Yuzhu Shi, Jian Liang, Wenxuan Zhang, Chao Xue, Vincent YF Tan, and Shaojun Bai. Understanding and mitigating dimensional collapse in federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2936–2949, 2023.

Jordan Shipard et al. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Patrick Von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Dolfi, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. *GitHub repository*, 2022.

Jianyu Wang et al. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in neural information processing systems*, 2020.

Tehrim Yoon et al. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv preprint arXiv:2107.00233*, 2021.

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

Guanyi Zhu, Xinyi Liu, Jianjun Niu, Shaojie Tang, Xiaofei Wu, and Jiqiang Zhang. Dualfed: Enjoying both generalization and personalization in federated learning via hierachical representations. In *ACM MM*, pp. 11060–11069, 2024.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

A Appendix

A.1 Experimental Setup

Datasets and Model We evaluate FedHDL on four benchmarks covering distinct non-IID characteristics to thoroughly verify its robustness. For label skew, we use CIFAR-10 with a Dirichlet Hsu et al. (2019) partition $\beta \in \{0.05, 0.1, 0.3, 0.5\}$ to simulate varying degrees of class imbalance. We partition CIFAR-10 among $K = 100$ clients, and in each communication round, we randomly sample a fraction of $C = 0.1$ (i.e., 10 clients) for local training. The classifier for CIFAR-10 Krizhevsky & Hinton (2009) is a CNN consisting of two convolutional layers. For feature skew, we extensively evaluate our method on three domain-partitioned benchmarks: **Office-Home** Venkateswara et al. (2017), **PACS** Li et al. (2017), and **DomainNet** Peng et al. (2019). For Office-Home, each of the $K = 4$ clients is assigned one specific domain (Art, Clipart, Product, or Real-world). Similarly, PACS distributes its $K = 4$ distinct domains (Photo, Art, Cartoon, and Sketch) across 4 clients, and DomainNet evaluates cross-domain scalability across its subset of domains. We utilize a ResNet-18 backbone He et al. (2016) for all feature-skewed tasks (Office-Home, PACS, and DomainNet). For all generative components, we adopt Stable Diffusion v1.5 with LoRA adapters of rank $r \in \{8, 16, 32, 64, 128\}$ to assess parameter efficiency.

Baselines Existing FL methods often target specific types of heterogeneity. Therefore, we select *setting-specific* representative baselines for each scenario to ensure a rigorous comparison. For the label-skewed CIFAR-10 (Dirichlet partition), we compare against a comprehensive suite of optimization and aggregation baselines, including FedAvg McMahan et al. (2017), FedProx Li et al. (2020), FedNova Wang et al. (2020), Scaffold Karimireddy et al. (2020), FedNTD Lee et al. (2022), FedSAM Qu et al. (2022), FedASAM Caldarola et al. (2022), and FedSOL Lee et al. (2024). These baselines cover various strategies from objective regularization to sharpness-aware minimization. For the feature-skewed settings (**Office-Home, PACS, and DomainNet**), we compare FedHDL against domain-adaptive and personalization strategies, including SingleSet (no federation), FedAvg McMahan et al. (2017), FedProx Li et al. (2020), FedPer Arivazhagan et al. (2019), FedRep Collins et al. (2021), FedBN Li et al. (2021), FedETF Li et al. (2023), and the strong hybrid baseline DualFed Zhu et al. (2024). Crucially, we include FedAvg and FedProx as common reference points across all scenarios to facilitate cross-table consistency and provide a clear understanding of the relative improvements achieved by our generative harmonization approach.

Implementation Details Across all settings, we use a Momentum SGD optimizer with a learning rate of 0.01, momentum of 0.9, and weight decay of 1×10^{-5}. All experiments are conducted for $T = 300$ communication rounds, and results are reported over three independent runs using random seeds $\{0, 1, 2\}$. On the label-skewed CIFAR-10 dataset, clients perform $E = 5$ local epochs per round with a batch size of 64. For the three domain-skewed datasets (Office-Home, PACS, and DomainNet), we set the local epoch to $E = 1$ with a batch size of 256 to ensure cross-domain training stability.

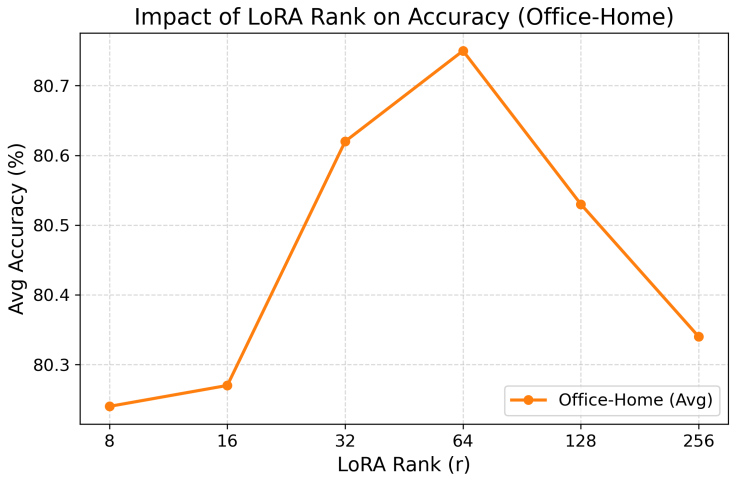


Figure 3: Impact of LoRA Rank (r) on the average accuracy of the Office-Home dataset. The performance peaks at $r = 64$, striking an optimal balance between model generalization (toward the left-hand side with small r) and stylistic representation (toward the right-hand side with large r).

Image Generation Settings For the generative harmonization process, we utilize the pre-trained `stable-diffusion-v1-5` checkpoint via the open-source `Diffusers` library Von Platen et al. (2022). To ensure high-fidelity and category-aligned sample generation, we employ 28 *inference steps* and a *guidance scale* of 7.5 during the synthesis stage across all four benchmarks. The images are generated using tailored prompt templates depending on the dataset characteristics. For the label-skewed CIFAR-10 dataset, we utilize the template: 'a photo of a {label}". For the domain-skewed datasets (Office-Home, PACS, and DomainNet), we employ a style-conditioned template: a photo of a {label} in {domain} style", where $\{label\}$ and $\{domain\}$ are dynamically filled based on the target class and the specific client's domain context, respectively.

A.2 Impact of LoRA Rank

We further investigate the sensitivity of FedHDL to the LoRA rank under both feature-skewed and label-skewed federated learning settings. The LoRA rank determines the dimensionality of the trainable low-rank subspace and therefore directly controls the adaptation capacity, parameter efficiency, and communication cost of the diffusion model. A smaller rank imposes a stronger low-dimensional constraint, which improves parameter efficiency and may provide an implicit regularization effect. In contrast, a larger rank increases the expressive capacity of the adapter, allowing it to capture more detailed semantic and stylistic variations from local client data. However, this additional flexibility may also increase the risk of learning client-specific noise or overly specialized local characteristics.

We analyze the effect of LoRA rank under both feature and label heterogeneity. On Office-Home, performance improves from $r = 8$ to $r = 64$, peaking at approximately 80.75%, and then slightly decreases for larger ranks. This suggests that $r = 64$ provides a good balance between adaptation capacity, cross-domain generalization, and parameter efficiency. Under label heterogeneity, varying $r \in \{8, 16, 32, 64, 128\}$ across different Dirichlet settings results in only modest performance changes, indicating that FedHDL is relatively robust to the choice of rank, although the optimal value may vary with the severity of non-IID data.

Table 6: Impact of LoRA rank r on CIFAR-10 accuracy under different label skew levels β. All results are reported as the average of 3 independent runs (mean $\pm$ std). **Bold** indicates the best performance for each setting.

LoRA Rank r	$\beta = 0.05$	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$
8	71.21 ± 0.45	71.44 ± 0.44	73.25 ± 0.69	73.27 ± 0.79
16	70.78 ± 0.67	71.32 ± 0.44	73.59 ± 0.80	73.14 ± 0.74
32	$\mathbf{71.41 \pm 0.76}$	71.51 ± 0.51	73.30 ± 0.88	$\mathbf{73.36 \pm 0.43}$
64	71.35 ± 0.50	$\mathbf{71.57 \pm 0.56}$	73.50 ± 0.70	73.20 ± 0.89
128	71.39 ± 0.48	71.50 ± 0.12	$\mathbf{73.64 \pm 0.55}$	73.33 ± 0.57