Out-of-Distribution Federated Distillation with Domain-Aware Proxy

Anonymous ACL submission

Abstract

Federated Learning is a distributed machine learning paradigm that trains a global model by aggregating local clients without sharing private data of each client. Federated Distillation (FD) builds upon this paradigm by leveraging knowledge distillation to exchange soft predictions on proxy data instead of model parameters, enabling more efficient communication and supporting heterogeneous model collaboration. However, FD models trained on In-Distribution data are hardly adapted to Out-of-Distribution (OOD) scenarios. In this paper, we propose a domain-aware proxy selection framework to better adopt proxy data for OOD problems. The experimental results show that the proposed models effectively address the challenges of distribution shifts under OOD with and without proxy data by achieving average 82.9% and 80.6% over existing works on standard benchmarks. The codes and data are released in https://anonymous.4open.science/r/DPS-FD-8596/.

1 Introduction

Federated Learning is an emerging distributed machine learning paradigm that trains a global model by aggregating local clients without sharing private data kept by each client (McMahan et al., 2017; Abadeer et al., 2022; Pei et al., 2024; Chen et al., 2024; Wang et al., 2025a). However, standard Federated Learning has the limitations of the high communication cost and the homogeneous customization. To address these limitations, Federated Distillation (FD) introduces shared proxy data as a medium for knowledge transfer between clients and the server (Hinton et al., 2015; You et al., 2017; Li and Wang, 2019). By exchanging soft labels on proxy data instead of model parameters, FD significantly reduces communication overhead and enables the use of heterogeneous models across clients (Lin et al., 2020; Hu et al., 2021; Wu et al., 2022; Itahara et al., 2023; Shao et al., 2024; Fan et al., 2025; Wang et al., 2025c).

However, FD models trained on In-Distribution data are hardly adapted to Out-of-Distribution (OOD) scenarios, due to the proxy data that serves as a critical medium for knowledge transfer (Hendrycks and Gimpel, 2016; Guo et al., 2023; Bai et al., 2023; Qi et al., 2024, 2025; Jeong and Choi, 2025). Since proxy data heavily affect distribution shifts under OOD, we propose **DPS-FD**, a **D**omain-aware **P**roxy **S**election framework for **FD** to better adopt proxy data for OOD problems. Specifically, DPS-FD enables each local client to select the relevant domain samples, while the server selects globally-representative samples from the proxy data. By taking the union of these selected samples, DPS-FD constructs a domain-aware proxy data that better captures the characteristics of both local and global distributions.

Although the effective adoption of proxy data can mitigate OOD promblem in FD, the original proxy data sometimes cannot be obtained for knowledge transfer in FD (Zhou and Chiam, 2023; Zhu, 2024; Fang et al., 2024). These works mainly focus on computer visions (Takahashi et al., 2023; Liao et al., 2024b; Qi et al., 2025; Wang et al., 2025b). To address the challenge of no proxy data in textual data, we propose a **V**ocabulary-**C**onstrained LLM-based generation (**VC**) strategy for generating proxy data, where a global vocabulary is adopted as lexical constraints to guide the proxy data generation. VC adopts a few-shot prompting strategy that consists of two complementary prompt templates: system prompts and user prompts, to generate diverse and high-quality proxy data. According to the comprehensive empirical experiments, we find that high-quality proxy data can significantly mitigate OOD problems. DPS-FD generates proxy data using the VC strategy if the proxy data is unavailable.

Experiment results on standard benchmarks

demonstrate that DPS-FD achieves competitive performance both with and without proxy data. Additionally, we take insight analysis on the effect of the high-quality diverse proxy data, unveiling that global distribution of the proxy data heavily affect the OOD performance of the global model in server. The main contribution of this paper are as follows:

- We propose a domain-aware proxy selection framework that better adopts proxy data for OOD problems, enhancing the robustness and generalization of the global model.
- We introduce a vocabulary-constrained LLM-based generation strategy to enable FD models to maintain competitive performance without access to real proxy data.
- We conduct analysis on the role of proxy data in FD, revealing that high-quality proxy data are crucial for mitigating OOD problems and enhancing the robustness of the global model.

2 Related Work

2.1 Federated Learning and Distillation

FL emerges as a promising distributed learning paradigm that enables collaborative model training without sharing private data of each client. Classic approaches such as FedAvg (McMahan et al., 2017) aggregate model parameters from clients into a global model and iteratively repeats this process. However, they suffer from high communication overhead, require homogeneous model architectures, and privacy-leakage across clients and the server (Sui et al., 2020; Lin et al., 2021). To mitigate these limitations, FD leverages knowledge distillation (Hinton et al., 2015; You et al., 2017; Anil et al., 2018) to exchange soft labels on shared proxy data instead of model parameters (Wu et al., 2022; Itahara et al., 2023; Chen et al., 2024; Wang et al., 2025c), significantly reducing communication costs, and enabling heterogeneous models.

2.2 OOD in Federated Distillation

OOD behavior manifests primarily as covariate shifts and semantic shifts (Liao et al., 2024b), which often leads to model degradation in FD (Gulrajani and Lopez-Paz, 2020). To address this issue, FD extends to OOD scenarios by enhancing the alignment between proxy and client distributions (Yu et al., 2023; Qi et al., 2025). Existing work leverages public or synthesized proxy data to reduce domain gaps and improve knowledge transfer (Jeong et al., 2023). Zhu et al. (2021) introduce adaptive weighting and ensemble strategies to emphasize client-specific contributions during distillation and enhance robustness. These developments highlight FD potential in improving generalization in OOD environments. To further advance this line of research, we propose DPS-FD, which constructs high-quality proxy data to model the diversity across clients, enabling effective and efficient knowledge transfer.

2.3 Proxy Data in Federated Distillation

Proxy data plays a crucial role in FD, serving as the essential medium through which knowledge is exchanged between heterogeneous clients and the global model (Li and Wang, 2019; Lin et al., 2020). Without proxy data, the distillation process struggles to align knowledge (Liao et al., 2023, 2024a; Xiao and Liu, 2025).

To address these limitations, existing studies explore various generative approaches to synthesize proxy data for FD. These include 1) logit-based inversion methods, which reconstruct pseudo data by optimizing inputs to match client logits (Takahashi et al., 2023); 2) generator-based approaches using GANs or VAEs to learn data distributions and produce representative samples (Zhang et al., 2022; Wang et al., 2023; Liao et al., 2024b; Ma et al., 2025; Qi et al., 2025); 3) diffusion-based models, which iteratively denoise random noise into high-quality synthetic data (Li et al., 2023; Wang et al., 2024; Yang et al., 2024; Wang et al., 2025b).

However, in natural language processing, the situation becomes more challenging. Most existing proxy-based distillation methods are designed for vision tasks and overlook fundamental linguistic properties such as lexical frequency distribution and semantic structure. This gap highlights the need for domain-aware proxy construction methods tailored to textual data and motivates our work on leveraging lexical constraints and large language models to build more effective proxy data in FD.

3 Preliminary

We take classification as an example to investigate OOD problems in FD. In the context of a classification task under FL with K clients and a central server, each client $k \in \{i = 1, 2, \ldots, K\}$ locally keeps its own private labeled data $D_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ that is isolated from others, where $x_i^k \in \mathbb{R}^d$ and $y_i^k \in \{1, 2, \ldots, C\}$ are the i-th in-

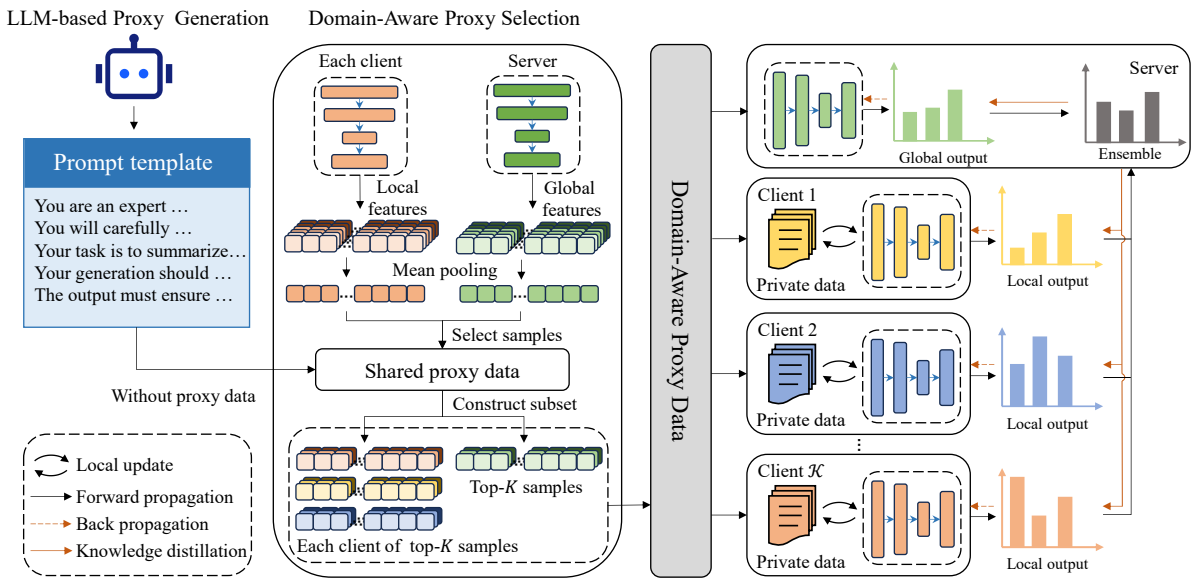


Figure 1: DPS-FD.

stance and its corresponding label, respectively, C is the number of classes, and N_k is the size of private data. The objective is to train a globally optimal model f_g parameterized by θ_g through aggregating client models f_k parameterized by θ_k in a decentralized manner without exposing private data on the server.

To mitigate the limitations of FL based on parameter exchange, FD leverages knowledge distillation to exchange soft labels instead of model parameters by introducing shared proxy data $D_p = \{(x_i^p)\}_{i=1}^{N_p}$ for knowledge transfer between clients and the server, where N_p is the size of the proxy data. Specifically, in each communication round $t \in T$, the server randomly selects the set of activated clients $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$ based on a sampling fraction ϵ to participate in the FD training, where $|\mathcal{K}| = |\epsilon \cdot K|$. Each client $\mathcal{K}$ first trains its local model on its own private data D_k, and the local optimization objective is defined as follows:

$$\theta_k^* = \arg\min_{\theta_k} \mathbb{E}_{(x,y)\sim D_k}[\mathcal{L}_{\text{CE}}(f_k(x;\theta_k), y)], \quad (1)$$

where $\mathcal{L}_{\text{CE}}(\cdot)$ denotes the cross-entropy loss used for local classification tasks, and $\mathbb{E}$ denotes its expected value with respect to the local data distribution D_k. The participating clients compute soft labels on the proxy data D_p and upload them to the server. The server then aggregates these labels using a weighting scheme proportional to the size of each client local data to distill the global model. The corresponding objective function is formulated as:

$$w_k = \frac{N_k}{||N||_1}, \; N = [N_1, N_2, \ldots, N_{\mathcal{K}}], \quad (2)$$

$$h(x_p) = \sigma(f(x_p;\theta)), \quad (3)$$

$$\theta_g^* = \arg\min_{\theta_g} \mathbb{E}_{x_p\sim D_p}\big[\mathcal{L}_{\text{KL}}\big(\sum_{k\in\mathcal{K}} w_k \cdot h_k(x_p),\; h_g(x_p)\big)\big], \quad (4)$$

where $\sigma(\cdot)$ is the softmax function with temperature τ to control the smoothness of soft labels, and $\mathcal{L}_{\text{KL}}$ is the Kullback-Leibler divergence used to measure the distribution differences between the server and the aggregated local models. Finally, each client further distills its local model using the KL divergence with the global soft labels predicted by the global model on the proxy data D_p.

$$\theta_k^* = \arg\min_{\theta_k} \mathbb{E}_{x_p\sim D_p}\big[\mathcal{L}_{\text{KL}}\big(h_g(x_p), h_k(x_p)\big)\big]. \quad (5)$$

4 Method

We construct OOD benchmarks in natural language processing tasks to investigate the challenges of FD in real-world scenarios. Then, we propose **DPS-FD**, a novel framework of **D**omain-aware **P**roxy **S**election for FD to address the OOD challenge by reconstructing the domain-aware proxy data. Furthermore, we introduce a vocabulary-constrained LLM-based proxy generation strategy coupled with the DPS mechanism to mitigate the OOD problems of FD without proxy data.

4.1 OOD Benchmark Construction

To better understand the challenges of FD in real-world scenarios, we investigate the role of proxy data in OOD settings. Before introducing the OOD setting, we first consider a multi-domain non-IID configuration in natural language processing, which accounts for both label distribution skew and the domain distribution of each client (Xiao and Liu, 2025; Yang et al., 2023; Xiao et al., 2024; Mao et al., 2025; Yan et al., 2025). Specifically, each client keeps private data from a distinct domain, and the label distribution across clients is made heterogeneous by sampling according to a Dirichlet distribution with different α parameters. Building upon the multi-domain non-IID setup, we simulate OOD settings by replacing the original test set, which is composed of data from the client domains, with a test set sampled from the domains not seen by the clients. Both the private data and proxy data are sampled from an open-source Amazon product review database.[1] The domain-level feature distributions are visualized in Appendix A.

4.2 Domain-Aware Proxy Selection

As shown in Figure 1, in communication round t, each client optimizes its local model using its labeled private data and utilizes the DPS mechanism to reconstruct domain-aware proxy data and uploads the corresponding local outputs to the server. The server aggregates the outputs and optimizes the global model through knowledge distillation. The global outputs given by the server are subsequently broadcast to all clients for local distillation, enabling bidirectional knowledge transfer.

DPS enables each client k to select the most relevant domain samples from the proxy data D_p, allowing clients to transmit high-confidence, domain-specific knowledge that enhances the robustness of the global model, while the server selects globally representative samples to improve the generalization of the global model. By taking the union of these selected samples, we construct a domain-aware proxy data D_p^* that serves as a more effective and efficient knowledge transfer medium between clients and the server.

In each communication round $t \in \{1, \ldots, T\}$, after local training on its private data D_k, client k computes the centroid representation c_k of its domain, and the server computes the global centroid c_g from the proxy data D_P as:

$$\begin{aligned} \mathbf{c}_k &= \frac{1}{N_k} \sum_{(x,y)\in D_k} f_k(x;\theta_k), \\ \mathbf{c}_g &= \frac{1}{N_p} \sum_{x\in D_p} f_g(x;\theta_g), \end{aligned} \tag{6}$$

where $f_k(x)$ and $f_g(x)$ denote the input x feature representations of the local model and the global model, respectively. For each proxy sample $x_i \in D_P$, we calculate its cosine similarity with each client centroid and the global centroid:

$$\begin{aligned} \mathrm{sim}_k(x_i) &= \frac{f_k(x_i;\theta_k)\cdot \mathbf{c}_k}{\|f_k(x_i;\theta_k)\|\,\|\mathbf{c}_k\|}, \\ \mathrm{sim}_g(x_i) &= \frac{f_g(x_i;\theta_g)\cdot \mathbf{c}_g}{\|f_g(x_i;\theta_g)\|\,\|\mathbf{c}_g\|}. \end{aligned} \tag{7}$$

Each client and the server then select the top-K proxy samples most similar to its domain representations, respectively:

$$\begin{aligned} P_k &= \mathrm{TopK}\big(\{x_i \in D_P \mid \mathrm{sim}_k(x_i)\}\big), \\ P_g &= \mathrm{TopK}\big(\{x_i \in D_P \mid \mathrm{sim}_g(x_i)\}\big). \end{aligned} \tag{8}$$

Finally, the domain-aware proxy data used for knowledge distillation is constructed as the union:

$$D_p^* = \bigcup_{k=1}^{\mathcal{K}} P_k \cup P_g. \tag{9}$$

This selection process is repeated in each communication round, ensuring that the proxy data dynamically aligns with evolving domain representations as the training progresses. DPS-FD algorithm is given in Appendix B.

4.3 LLM-Based Proxy Generation

The global model benefits from additional knowledge that helps align the client-specific distributions and facilitates more effective knowledge transfer. High-quality proxy data provides a bridge for the global model to capture patterns not fully represented in individual client private data, enhancing both robustness to domain-specific variations and generalization to unseen domains. In the absence of proxy data, the global model has to rely solely on local clients, leading to weak alignment among local distributions.

To address the challenge of unavailable proxy data in FD, we propose a vocabulary-constrained LLM-based proxy data generation strategy. As

[1] https://archive.org/details/amazon-reviews-1995-2013.

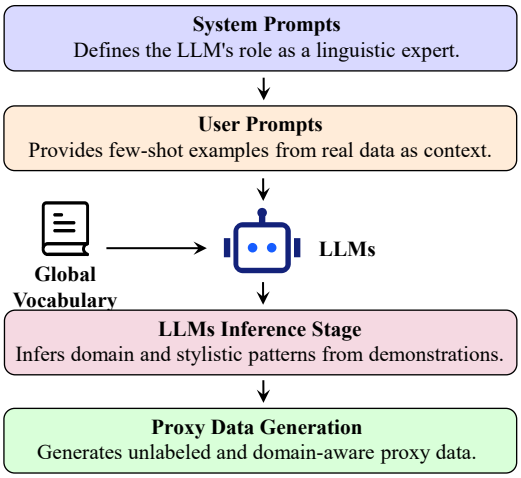


Figure 2: LLM-based Proxy Generation with Vocabulary-Constraints.

Model	g.	a.	b.	c.	h.	s.	cost (%)	ECE	entropy
with proxy data									
FD-LP	80.7	79.8	84.7	91.4	77.4	81.8	100	13.89	0.18
FD-GP	81.4	80.3	84.1	85.0	75.6	78.7	100	7.03	0.33
DPS-FD	**82.9**	80.6	85.1	81.9	76.0	78.9	98.3	9.84	0.20
without proxy data									
FD-VC	78.2	75.0	80.5	73.9	68.4	72.7	100	16.58	0.18
DPS-FD	**80.6**	77.5	81.7	74.6	73.5	76.3	95.9	8.68	0.29
w/o VC	77.7	75.8	79.3	77.1	66.6	71.6	93.9	7.00	0.40

Table 1: Results of the global model in heterogeneous OOD settings ($p < 0.001$). Column **g.** is score on test data in global distribution, while columns **a.**, **b.**, **c.**, **h.** and **s.** are scores on test data in respective domains.

shown in Figure 2, we first construct a vocabulary from tokenizing real proxy data, which serves as a constraint to ensure that generated samples adhere to realistic lexical distributions and domain-specific semantics. To effectively generate diverse and high-quality proxy data, we adopt a few-shot prompting strategy, consisting of two prompts:

- System prompts define the overall role and objective of the LLM, instructing it to act as a linguistic expert specialized in generating diverse, fluent, and contextually coherent product review texts tailored for FL tasks.
- User prompts provide few-shot representative samples drawn from real environments as in-context demonstrations.

To ensure linguistic consistency and data complexity, we constrain the average sentence length to approximately 90 words per generated text. As shown in Figure 2, two-stage prompting design allows the LLM to capture both global generation intent and domain-specific semantics. [2]

5 Experiments

We evaluate our proposed methods, DPS-FD along with a vocabulary-constrained LLM-based generation under OOD both with and without proxy data.

5.1 Settings

We conduct experiments on a sentiment classification task under heterogeneous client settings, considering both the presence and absence of proxy data. To simulate OOD scenarios, data are sampled by label within each domain from the Dirichlet distribution with concentration parameters $\alpha = \{0.3, 0.5, 0.8, 1.2, 1.5\}$ for clients, respectively, to control the degree of heterogeneity.

Backbones We assign different pre-trained language models to the 5 clients, namely `BERT-base-cased`, `BERT-large-cased`, `RoBERTa-base`, `RoBERTa-large`, and `XLNet-large-cased` model, respectively, while using `RoBERTa-large` as the server model.[3]

Data Clients keep local data from specific domains, i.e., automotive (a.), baby (b.), clothing (c.), health (h.), and sport (s.), respectively, while the global distribution (g.) consists of all domains unseen by the clients.

Metrics We use F1 scores to evaluate classification performance, cost to measure communication efficiency during knowledge transfer, and ECE together with prediction entropy to assess model calibration and uncertainty analysis.

5.2 Training and Inference

We train the models for 5 communication rounds with 3 local epochs per round, using an initial learning rate of 2e-5. AdamW is adopted as the optimizer, with a maximum sequence length of 128 and a batch size of 32. The number of clients is set to 5, corresponding to different numbers of domain-specific data. For the proposed DPS-FD method, we select the top 25k samples on each client and the top 35k samples on the server to form the refined proxy data. GPT-3.5-turbo is used as the LLM in the LLM-based generation strategy. We use an NVIDIA GPU A100 for training and inference.

[2] The detailed structures of the system and user prompt are given in Appendix C.

[3] All the pre-trained language model cards are sourced from `https://huggingface.co/models`.

Model	a.	b.	c.	h.	s.
with proxy data					
FD-LP	78.2	47.6	93.4	76.9	81.5
FD-GP	78.0	47.6	93.2	76.0	81.4
DPS-FD	78.5	84.6	90.4	76.5	82.3
without proxy data					
FD-VC	76.3	47.6	94.0	73.3	77.7
DPS-FD	76.9	81.0	89.8	67.7	77.1
w/o VC	75.2	82.1	91.3	70.0	72.8

Table 2: Results of local models on clients ($p < 0.001$). The columns **a.**, **b.**, **c.**, **h.** and **s.** are scores on test data in respective domains.

DPS-FD	top-K (client, server)	g.	cost (%)	ECE	entropy
5 Clients					
w/ proxy data	(15k, 15k)	82.6	89.8	9.68	0.22
	(25k, 35k)	**82.9**	98.3	9.84	0.20
w/o proxy data	(15k, 15k)	77.5	78.3	13.68	0.26
	(25k, 35k)	80.6	96.3	8.68	0.29
10 Clients					
w/ proxy data	(15k, 15k)	**84.2**	95.9	7.36	0.22
	(25k, 35k)	84.0	99.8	11.69	0.13
w/o proxy data	(15k, 15k)	81.4	80.3	10.97	0.20
	(25k, 35k)	82.7	93.6	5.86	0.28

Table 3: Results of DPS-FD across different number of clients and top-K selections.

5.3 Baselines

We evaluate the proposed DPS-FD against a series of baselines to thoroughly investigate the impact of proxy data and the effectiveness of our approach under OOD settings.

- **FD-LP** refers to the standard FD method (Lin et al., 2020), which uses proxy data consisting only of samples from the client domains.
- **FD-GP** uses global proxy data composed of samples from all domains that are unseen by the clients.
- **FD-VC** leverages a vocabulary-constrained LLM-based generation strategy to synthesize proxy data without proxy data.
- **DPS-FD** is our proposed model, applying DPS on FD with global domain proxy data, with proxy data generated by a vocabulary-constrained LLM, and with proxy data generated by an unconstrained LLM.

5.4 Results

Table 1 and 2 show the results of the global server model and the local client models, respectively, on our proposed model and several baseline models.

Proxy Data Influence As shown in Table 1, the global model of FD-GP outperforms FD-LP, indicating that global-domain proxy data, compared to local-domain proxy data, can more effectively mitigate distribution shifts under OOD settings.

With proxy data, our proposed DPS-FD achieves superior performance across both global and local domains except for the clothing domain. By incorporating DPS, clients and the server dynamically identify and utilize proxy samples that better align with the distributions of local and global domains. DPS effectively filters out noisy or irrelevant samples, leading to high-quality proxy data that enable more effective and efficient knowledge transfer. Without proxy data, DPS-FD outperforms DPS-FD without vocabulary-constraint by approximately 2.9%, demonstrating that vocabulary-constrained LLM generation produces higher-quality proxy data. Moreover, DPS-FD substantially outperforms FD-VC, remaining effective even with synthetic proxy data.

Calibration and Uncertainty Estimation As shown in Table 1, with proxy data, replacing local-domain proxy data (FD-LP) with global-domain proxy data (FD-GP) significantly improves robustness under distribution shifts, resulting in lower ECE and higher entropy. Although DPS-FD does not achieve the lowest ECE, it maintains balanced calibration with moderate entropy, avoiding overconfident predictions. Without proxy data, FD-VC exhibits severe overconfidence, while incorporating DPS improves robustness with lower ECE and higher entropy. Although DPS-FD without VC achieves the lowest ECE, the corresponding performance drop suggests that excessive diversity may introduce noisy samples, suggesting the benefit of vocabulary-constrained generation for balancing performance and robustness.

Local Models Table 2 shows F1 scores of each local model evaluated on their own local test set. Existing methods perform poorly on local distribution of the baby domain, achieving only 47.6% F1 scores, whereas our proposed DPS-FD substantially achieves the improvements on the local test set. DPS effectively mitigates the interference from other domains and improves the performance of local models within their own domains by filtering out noise and irrelevant proxy samples. This demonstrates that DPS-FD not only enhances the global model on server but also somehow enhances the local model on clients.

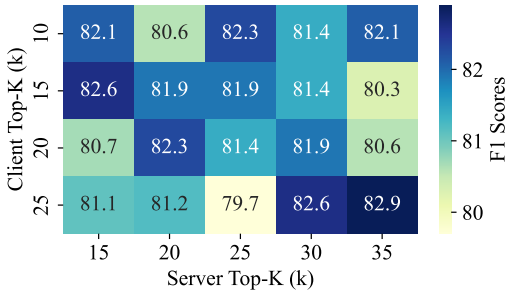


Figure 3: F1 scores varies with the top-K selections of the clients and server

Model	g.	a.	b.	c.	h.	s.	cost (%)	ECE	entropy
DS-FL	80.1	79.7	82.0	78.9	78.7	80.3	100	16.74	0.12
MHAT	81.6	79.5	83.6	80.3	72.6	78.6	100	1.66	0.44
FedKD	82.0	80.0	84.1	82.8	73.2	78.5	16.7	2.80	0.41
DPS-FD	**82.9**	80.6	85.1	81.9	76.0	78.9	98.3	9.84	0.20

Table 4: Results of the global model for existing FD methods under heterogeneous and OOD settings where proxy data is provided.

Communication Cost To facilitate a fair comparison, we normalize the communication cost of the traditional FD method to 100% as a baseline. As shown in the column cost of Table 1, DPS-FD achieves a slight yet meaningful reduction in communication cost from 100% to 98.3% with proxy data and from 100% to 95.9% without proxy data. The overall cost is determined by the number of clients $\mathcal{K}$, the proxy data size $|D_P|$, and the number of communication rounds T, and is typically expressed as $(\mathcal{K}+1)\cdot|D_P|\cdot T$. By constructing a domain-aware proxy data $|D_P^*|$, DPS-FD improves model performance and reduces unnecessary communication by selecting representative samples that capture the distribution of each domain.

5.5 Comparisons with SOTA Models

As shown in Table 4, our proposed DPS-FD significantly outperforms existing FD methods, including DS-FL (Itahara et al., 2023), MHAT (Hu et al., 2021), and FedKD (Wu et al., 2022), on F1 scores under global distribution performance with proxy data. In terms of calibration and uncertainty estimation, DS-FL exhibits poor calibration performance with very low entropy, indicating severe overconfidence in its predictions. In contrast, MHAT and FedKD achieve lower ECE values but produce higher entropy, suggesting more conservative predictions. Our proposed DPS-FD maintains moderate calibration performance with balanced entropy levels, avoiding both excessive overconfidence and overly conservative behavior. Moreover, DPS-FD achieves this improved performance while reducing communication costs compared to DS-FL and MHAT, benefiting from an effective knowledge exchange process, except for FedKD, which adopts a one-shot distillation paradigm. This demonstrates superior robustness and generalization across heterogeneous clients, highlighting the effectiveness of domain-aware proxy selection in mitigating distribution shift and improving reliability in FD.

6 Discussion and Analysis

We analyze the performance of DPS-FD with respect to the number of clients, top-K selection, proxy data quality, and computational overhead.

6.1 The Number of Clients

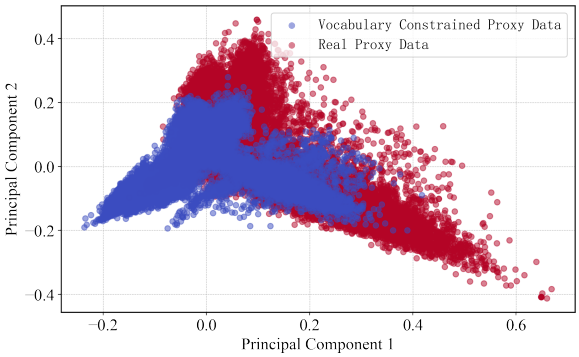
To investigate the effect of client number, five additional clients from the beauty, patio, pet, shoes, and software domains of the Amazon data are introduced, using `BERT-base-cased` as the local model.

As shown in Table 3, global performance consistently improves as the number of participating clients increases, both with and without proxy data. This trend suggests that additional clients provide more diverse domain information, enabling the global model to learn more robust representations under distribution shifts. Meanwhile, increasing the number of clients generally reduces ECE with slightly lower entropy, indicating more reliable predictions. An exception is observed under the w/ proxy setting with top-K (25k, 35k), where ECE increases when scaling from 5 to 10 clients. This may be due to the enlarged proxy set introducing less informative samples, which slightly degrades calibration reliability.

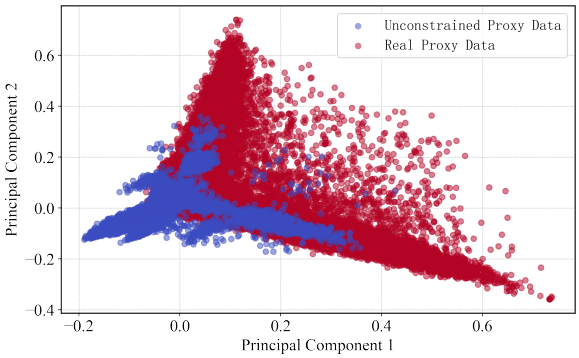
6.2 Top-K Selection

In DPS-FD, the choice of the top-K samples selected by the client and the server plays a crucial role, as it directly affects the trade-off between model performance and communication cost. We vary the number of selected samples on the client side as {10k, 15k, 20k, 25k} and on the server side as {15k, 20k, 25k, 30k, 35k} under the setting with proxy data.

As shown in Table 3, with proxy data, the (15k, 15k) configuration achieves a substantial reduction in communication cost (approximately 8.5% and 3.9%) while maintaining comparable global performance. Without proxy data, the same (15k, 15k)



(a) Vocabulary-constrained proxy data



(b) Unconstrained proxy data

Figure 4: Word distributions of proxy data generated under vocabulary-constrained (top) and unconstrained (bottom) settings, represented in the TF-IDF reduced dimensionality space.

configuration yields even larger reductions of 18% and 13.3%, respectively. As shown in Figure 3, performance does not monotonically improve with smaller top-K values. Smaller top-K retains only highly domain-relevant samples, aggressively filtering out most hard samples, which may remove informative near-boundary instances and limit generalization. In contrast, larger top-K preserves a portion of such hard samples, which helps improve generalization and decision boundary learning. These results suggest that retaining an appropriate proportion of hard samples is beneficial, rather than strictly filtering them out.

6.3 Quality of Proxy Data

We take the insight analysis on vocabulary constraints for the quality of proxy data generated by LLMs. We compare two generation strategies: with vocabulary constraints and without constraints. We tokenize the real data to construct a global vocabulary and enforce the LLM to generate samples using only the words in vocabulary. The unconstrained LLM directly generates samples based on same prompts without vocabulary constraints.

By constraining the vocabulary during LLM-based generation, we guide lexical choices toward distributions closer to real-world data. This reduces spurious lexical variations and improves semantic consistency of generated samples. As shown in Figure 4 (top), proxy data generated with vocabulary constraints is significantly closer to real proxy data, demonstrating the effectiveness of the proposed constraint mechanism. In contrast, unconstrained generation relies solely on prompt instructions, leading to more diverse but less controlled outputs. As shown in Figure 4 (bottom), such samples deviate from real proxy distributions, introducing irrelevant lexical patterns.

Together with the results in Table 1, these findings indicate that vocabulary constraints help ensure that proxy data remain semantically aligned and more suitable for distillation.

6.4 Computational overhead

DPS-FD introduces a lightweight proxy selection step performed locally at each client and the server. The computational complexity is $\mathcal{O}((\mathcal{K}+1)\cdot|D_P|\cdot d)$, where d is the embedding dimension. Compared to local training and distillation, which require full forward and backward passes of deep models, the proxy selection only involves cosine similarity computations and thus introduces negligible overhead. Therefore, the size of the proxy data does not become a bottleneck for the scalability of DPS-FD. The LLM-based proxy data generation is a one-time offline process and does not affect per-round training or communication cost.

7 Conclusion

In this paper, we propose domain-aware proxy selection that constructs a representative proxy data by jointly considering client-specific and global domain samples to better adopt the proxy data for OOD problems in FD. Additionally, to address the absence of proxy data, we introduce a vocabulary-constrained LLM-based proxy data generation strategy, which mitigates the generation of linguistically plausible but semantically irrelevant or inconsistent samples. By incorporating the proxy selection strategy with the generation strategy, DPS-FD enhances the robustness and generalization of FD both with and without proxy data. Our experimental results demonstrate that our models outperform existing methods under OOD settings.

Limitations

Our method relies on LLMs to synthesize proxy data. LLM-generated data are less reproducible and incur higher API costs and longer response time. We only simulate 5 and 10 clients in our experiments, and we believe that using more clients would be more effective. We take classification as a representative example in this work, and need to conduct more comprehensive validation. We will further explore strategies to reduce ECE while maintaining reasonable uncertainty levels and performance.

References

Macarious Abadeer, Wei Shi, and Jean-Pierre Corriveau. 2022. Flightner: A federated learning approach to lightweight named-entity recognition. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 687–694. IEEE.

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*.

Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. 2023. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, pages 1454–1471. PMLR.

Yuhang Chen, Wenke Huang, and Mang Ye. 2024. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12077–12086.

Tao Fan, Hanlin Gu, Xuemei Cao, Chee Seng Chan, Qian Chen, Yiqiang Chen, Yihui Feng, Yang Gu, Jiaxiang Geng, Bing Luo, and 1 others. 2025. Ten challenging problems in federated foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. 2024. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1266.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. 2023. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11905–11933. PMLR.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Li Hu, Hongyang Yan, Lang Li, Zijie Pan, Xiaozhang Liu, and Zulong Zhang. 2021. Mhat: An efficient model-heterogenous aggregation training scheme for federated learning. *Information Sciences*, 560:493–503.

Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. 2023. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1):191–205.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. 2023. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *Preprint*, arXiv:1811.11479.

Gu-Bon Jeong and Dong-Wan Choi. 2025. Out-of-distribution detection via outlier exposure in federated learning. *Neural Networks*, 185:107141.

Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. 2023. Is synthetic data from diffusion models ready for knowledge distillation? *Preprint*, arXiv:2305.12954.

Xinting Liao, Chaochao Chen, Weiming Liu, Pengyang Zhou, Huabin Zhu, Shuheng Shen, Weiqiang Wang, Mengling Hu, Yanchao Tan, and Xiaolin Zheng. 2023. Joint local relational augmentation and global nash equilibrium for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1536–1545.

Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Fengyuan Yu, Huabin Zhu, Binhui Yao, Tao Wang, Xiaolin Zheng, and Yanchao Tan. 2024a. Rethinking the representation in federated unsupervised learning with non-iid data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22841–22850.

Xinting Liao, Weiming Liu, Pengyang Zhou, Fengyuan Yu, Jiahe Xu, Jun Wang, Wenjie Wang, Chaochao Chen, and Xiaolin Zheng. 2024b. Foogd: Federated

collaboration for both out-of-distribution generalization and detection. *Advances in Neural Information Processing Systems*, 37:132908–132945.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363.

Tao Lin, Lingjing Kong, Sebastian U. Stich, and Martin Jaggi. 2021. Ensemble distillation for robust model fusion in federated learning. *Preprint*, arXiv:2006.07242.

Xinge Ma, Jin Wang, and Xuejie Zhang. 2025. Data-free black-box federated learning via zeroth-order gradient estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19314–19322.

Wenjie Mao, Bin Yu, Chen Zhang, AK Qin, and Yu Xie. 2025. Fedkt: Federated learning with knowledge transfer for non-iid data. *Pattern Recognition*, 159:111143.

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

Jiaming Pei, Wenxuan Liu, Jinhai Li, Lukun Wang, and Chao Liu. 2024. A review of federated learning methods in heterogeneous scenarios. *IEEE Transactions on Consumer Electronics*, 70(3):5983–5999.

Zhuang Qi, Weihao He, Xiangxu Meng, and Lei Meng. 2024. Attentive modeling and distillation for out-of-distribution generalization of federated learning. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Zhuang Qi, Sijin Zhou, Lei Meng, Han Hu, Han Yu, and Xiangxu Meng. 2025. Federated deconfounding and debiasing learning for out-of-distribution generalization. *arXiv preprint arXiv:2505.04979*.

Jiawei Shao, Fangzhao Wu, and Jun Zhang. 2024. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*, 15(1):349.

Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. FedED: Federated learning via ensemble distillation for medical relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2128, Online. Association for Computational Linguistics.

Hideaki Takahashi, Jingjing Liu, and Yang Liu. 2023. Breaching fedmd: image recovery via paired-logits inversion attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12198–12207.

Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. 2023. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20412–20421.

Naibo Wang, Yuchen Deng, Wenjie Feng, Jianwei Yin, and See-Kiong Ng. 2024. Data-free federated class incremental learning with diffusion-based generative memory. *Preprint*, arXiv:2405.17457.

Zheng Wang, Zihui Wang, Zheng Wang, Xiaoliang Fan, and Cheng Wang. 2025a. Federated learning with domain shift eraser. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4978–4987.

Zhongwei Wang, Tong Wu, Zhiyong Chen, Liang Qian, Yin Xu, and Meixia Tao. 2025b. Diffusion model-based data synthesis aided federated semi-supervised learning. *Preprint*, arXiv:2501.02219.

Zichen Wang, Feng Yan, Tianyi Wang, Cong Wang, Yuanchao Shu, Peng Cheng, and Jiming Chen. 2025c. Fed-dfa: Federated distillation for heterogeneous model fusion through the adversarial lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21429–21437.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. 2022. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1).

Canran Xiao and 1 others. 2024. Confusion-resistant federated learning via diffusion-based data harmonization on non-iid data. *Advances in Neural Information Processing Systems*, 37:137495–137520.

Jiahao Xiao and Jiangming Liu. 2025. Adaptive federated distillation for multi-domain non-iid textual data. *arXiv preprint arXiv:2508.20557*.

Guochen Yan, Luyuan Xie, Xinyi Gao, Wentao Zhang, Qingni Shen, Yuejian Fang, and Zhonghai Wu. 2025. 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, pages 21904–21912.

Lei Yang, Jiaming Huang, Wanyu Lin, and Jiannong Cao. 2023. Personalized federated learning on non-iid data via group-based meta-learning. *ACM Transactions on Knowledge Discovery from Data*, 17(4):1–20.

Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. 2024. Feddeo: Description-enhanced one-shot federated learning with diffusion models. *Preprint*, arXiv:2407.19953.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international*

conference on knowledge discovery and data mining, pages 1285–1294.

Shuyang Yu, Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. 2023. Turning the curse of heterogeneity in federated learning into a blessing for out-of-distribution detection. In *2023 International Conference on Learning Representations*.

Lan Zhang, Dapeng Wu, and Xiaoyong Yuan. 2022. Fedzkt: Zero-shot knowledge transfer towards resource-constrained federated learning with heterogeneous on-device models. In *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, pages 928–938. IEEE.

Tianxun Zhou and Keng-Hwee Chiam. 2023. Synthetic data generation method for data-free knowledge distillation in regression neural networks. *Expert Systems with Applications*, 227:120327.

Jun Zhu. 2024. Synthetic data generation by diffusion models. *National Science Review*, 11(8):nwae276.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR.

A Domain Diversity

To verify that the selected domains exhibit meaningful distribution gaps, we visualize the feature distributions of different domains using Principal Component Analysis (PCA). As shown in Figure 5, samples from different domains show clear separation, indicating non-trivial distribution shifts among domains. This supports the suitability of the Amazon Reviews dataset for evaluating domain shift and OOD generalization.

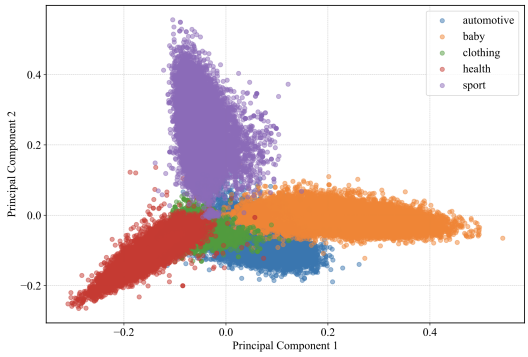


Figure 5: Domain distributions across different domains.

B Algorithm of Domain-Aware Proxy Selection

Algorithm 1 summarizes the overall workflow of DPS-FD and details how the proposed domain-aware proxy selection is incorporated into the FD process.

System Prompt Template

You are an expert language model specializing in generating diverse, high-quality product review texts for federated learning.
You will carefully analyze a set of example reviews and identify corresponding domains and sentiment tendencies.
While the examples may be biased toward certain domains, your task is to summarize them and infer additional potential domains.
Your generation should reflect the authentic style and tone of product reviews: natural, varied, and customer-oriented.
The output must ensure both domain diversity and linguistic diversity, avoiding repetitive templates while maintaining realism.

Algorithm 1: Domain-Aware Proxy Selection

Input: labeled private data $\{D_k\}_{k=1}^{\mathcal{K}}$; unlabeled proxy data D_p; global model θ_g; local models $\{\theta_k\}_{k=1}^{\mathcal{K}}$; communication rounds T
Output: global model θ_g
if D_p **then**
  Use the real proxy data D_p ;
else
  Generate a synthetic proxy data D_p using LLM-based generation strategy ;
end
for *each communication round* $t = 1, 2 \ldots, T$ **do**
  Client executes:
  for *each client* k *in parallel* **do**
    Train local model via **Eq.** (1) ;
    Select and upload the domain-specific proxy data via **Eq.** (6)(7)(8) ;
  end
  Server executes:
  Select the global domain proxy data via **Eq.** (6)(7)(8) ;
  Construct a domain-aware proxy data D_p^* via **Eq.** (9) ;
  Update the global model on D_p^* via **Eq.** (4) ;
  Client executes:
  for *each client* k *in parallel* **do**
    Update the local model on D_p^* via **Eq.** (5) ;
  end
end

C Prompt Templates

The following prompt templates are used to guide the LLM in generating proxy data. We design both user and system prompts to control the generation process, where the user prompt specifies generation constraints and the system prompt defines overall roles. These prompts are specifically designed to ensure that the synthesized proxy data can better approximate real-world review distributions under different domains in federated learning settings.

User Prompt Template

A set of representative examples.
Your task:
1. Identify the product domains and sentiment tendencies represented in the examples above, and infer additional potential domains that could reasonably exist.
2. Based on the examples' and inferred domains, generate exactly realistic and diverse unlabeled review texts.
3. Each review should be fluent, reflecting the natural style of customer reviews.
4. Each review should be about 90 words (minimum 40, maximum 180). Do not shorten because you need to output many.
5. The reviews should not be templates or mechanically repeated. Maintain variability in tone, sentence structure, and vocabulary.
Output format:
1. Output each reviews on a separate line.
2. Ensure the quality of each generated sentence and do not sacrifice quality for the sake of quantity.
3. Do not add any prefix, numbering, headers, or label—only the raw reviews.
Begin now: