Federated Deconfounding and Debiasing Learning for Out-of-Distribution Generalization

Zhuang Qi[1], **Sijin Zhou**[2], **Lei Meng**[1*], **Han Hu**[3], **Han Yu**[4], **Xiangxu Meng**[1]

[1]School of Software, Shandong University, China
[2]AIM Lab, Faculty of Engineering, Monash University, Clayton, VIC, Australia
[3]School of information and Electronics, Beijing Institute of Technology, China
[4]College of Computing and Data Science, Nanyang Technological University, Singapore
z_qi@mail.sdu.edu.cn, sjzhou1995@gmail.com, lmeng@sdu.edu.cn, hhu@bit.edu.cn,
han.yu@ntu.edu.sg, mxx@sdu.edu.cn

Abstract

Attribute bias in federated learning (FL) typically leads local models to optimize inconsistently due to the learning of non-causal associations, resulting degraded performance. Existing methods either use data augmentation for increasing sample diversity or knowledge distillation for learning invariant representations to address this problem. However, they lack a comprehensive analysis of the inference paths, and the interference from confounding factors limits their performance. To address these limitations, we propose the <u>Fed</u>erated <u>D</u>econfounding and <u>D</u>ebiasing <u>L</u>earning (FedDDL) method. It constructs a structured causal graph to analyze the model inference process, and performs backdoor adjustment to eliminate confounding paths. Specifically, we design an intra-client deconfounding learning module for computer vision tasks to decouple background and objects, generating counterfactual samples that establish a connection between the background and any label, which stops the model from using the background to infer the label. Moreover, we design an inter-client debiasing learning module to construct causal prototypes to reduce the proportion of the background in prototype components. Notably, it bridges the gap between heterogeneous representations via causal prototypical regularization. Extensive experiments on 2 benchmarking datasets demonstrate that FedDDL significantly enhances the model capability to focus on main objects in unseen data, leading to 4.5% higher Top-1 Accuracy on average over 9 state-of-the-art existing methods.

1 Introduction

Federated out-of-distribution (OOD) generalization typically leverages data with diverse attributes across clients for collaborative modeling. The aim is to enhance model performance on unseen distributions [Qi and et al., 2024; Qi *et al.*, 2025b; Wang *et al.*, 2024; Fu *et al.*, 2025b]. It allows FL clients to
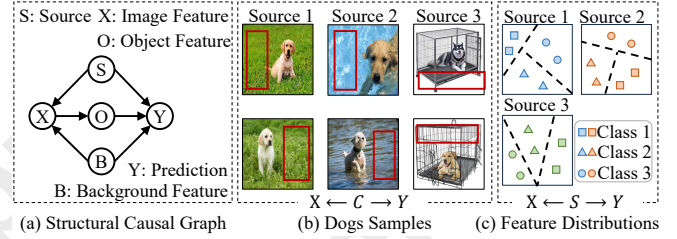
---

*Corresponding author

Figure 1: FedDDL reveals two factors affecting model inference: specific backgrounds and the output gap between data sources. (a) It constructs a structured causal graph to support the claim. (b) Dogs Samples illustrates the specific background within the client. (c) It represents the feature distribution differences between data sources.

train models locally and iteratively exchange parameters with the server, enabling global aggregation without exposing private data [Liao *et al.*, 2024; Fu *et al.*, 2025a; Hu *et al.*, 2023; Yi *et al.*, 2023]. Despite its contributions to data privacy protection, existing studies show that attribute skew among sources often leads to performance degradation in FL models [Fu *et al.*, 2025a; Zhang and et al., 2024; Zhang *et al.*, 2025a; Ren *et al.*, 2025]. This is primarily due to spurious correlations in local models, which hinder the formation of a robust global model during aggregation [Yang *et al.*, 2020; Hu *et al.*, 2024; Cai *et al.*, 2024b].

Existing approaches for enhancing FL model performance on OOD samples can be broadly divided into two groups: 1) knowledge distillation-based methods [Wei and Han, 2024; Fan *et al.*, 2025; Yu *et al.*, 2023; Huang *et al.*, 2025] and 2) data augmentation-based methods [Chen *et al.*, 2023; Shenaj *et al.*, 2023; Park *et al.*, 2024]. The former either treats the teacher's knowledge (e.g., the output of global or other local models) as a regularizer to provide additional supervision or decouples category-irrelevant features in the latent space, thereby mitigating the interference of irrelevant attributes. For instance, MCGDM [Wei and Han, 2024] shares classifiers across sources to match gradients both within and across domains to reduce overfitting to attribute features. DFL [Luo *et al.*, 2022] applies mutual information-based disentanglement to address negative transfer due to attribute skew. However, the suboptimal performance of the teacher model on OOD samples and the single environment limit the effects

of knowledge distillation and attribute decoupling. The latter approach typically trains generators or uses pre-trained generative models to create data with diverse attributes (e.g., CCST [Chen *et al.*, 2023]). For example, FedCCRL [Wang and Tang, 2024] fuses different samples for data augmentation. However, they often introduce privacy risks with performance limited by the quality of the generated data.

To address these limitation, we propose the Federated Deconfounding and Debiasing Learning (FedDDL) method. As shown in Figure 1, it focuses on the key variables involved in the data and model inference (including data sources, background, images, objects and labels) with a structural causal model (SCM). To mitigate the joint interference of data source and background factors on model inference, we designs two main modules: 1) the intra-client deconfounding learning (DEC) module, and 2) the inter-client debiasing learning (DEB) module. Specifically, DEC decouples the background and the object in images, generating counterfactual samples to establish the association between the background and all classes without sharing any sample information among FL clients. This ensures that the model does not erroneously interpret the background as a causal factor. DEB constructs causal prototypes using object images rather than original images, which reduces the proportion of specific attributes in the prototype components. In addition, it treats causal prototypes as templates to align representations across different sources. This encourages local models to focus on target objects rather than the background, thereby promoting the learning of a unified representation space across clients.

Extensive experiments were conducted on two datasets, including performance comparisons, ablation studies and case studies with visual attention visualizations to investigate the association between background and labels. The results demonstrate that FedDDL effectively mitigates the interference of background and focuses on the objects of samples in unseen cases, improving the effectiveness of collaborative learning. Compared to nine state-of-the-art existing methods, FedDDL achieves 4.5% higher Top-1 Accuracy on average.

2 Related Work

2.1 Data Augmentation-based Methods

To mitigate attribute bias, data augmentation-based methods [Wang and Tang, 2024; Shenaj *et al.*, 2023; Liu *et al.*, 2021; Xu *et al.*, 2023; Zhang *et al.*, 2024; de Luca and et al., 2022; Zhang *et al.*, 2025b; Qi *et al.*, 2022] have been developed to improve model generalization to previously unseen attributes by boosting the diversity of the data attributes. These methods typically rely on two approaches: 1) training data generators to create novel samples, or 2) utilizing pre-trained diffusion models to enrich sample diversity. The first approach involves exchanging local information between clients to generate new samples from different domains. For instance, FIST [Nguyen *et al.*, 2024] and StableFDG [Park *et al.*, 2024] produce samples with varying styles by sharing style-related information across clients. The second approach leverages prompt-driven techniques to generate samples that align with certain criteria [Morafah and et al., 2024; Zhao *et al.*, 2023]. Although enhancing diversity, these strate-

gies often introduce privacy concerns. Moreover, the quality differences between generated and original samples also limit their effectiveness.

2.2 Knowledge Distillation-based Methods

Knowledge distillation-based approaches utilize generalized knowledge to guide local models in extracting shared features that are independent of specific attributes. These methods typically adopt two main strategies: 1) regularizing representations from different sources to align them, thus promoting consistency in representations across diverse contexts [Wang *et al.*, 2023; Qi *et al.*, 2024; Qi *et al.*, 2025a; Qi *et al.*, 2023; Liu *et al.*, 2021; Sun *et al.*, 2023; Meng *et al.*, 2024; Zhu *et al.*, 2024]; or 2) decoupling invariant features within the latent space to minimize the impact of confounding factors [Yu *et al.*, 2023; Yi *et al.*, 2024; Yu and et al., 2011; Cai *et al.*, 2024a; Liu *et al.*, 2024]. For instance, FedProc [Mu *et al.*, 2023] and FPL [Huang *et al.*, 2023] leverage prototypes to enforce consistent representation learning across all clients, thereby encouraging them to converge within a shared representation space. MCGDM [Wei and Han, 2024] mitigates overfitting within individual domains by employing both intra-domain and cross-domain gradient matching to improve generalization. Similarly, FedIIR [Guo *et al.*, 2023] enables the model to implicitly learn invariant relationships by capitalizing on prediction inconsistencies and gradient alignment across clients. Although these techniques have achieved notable performance improvements, they often rely on single-domain data, which hinders model transferability to OOD domains. Our FedDDL method bridges these gaps.

3 Preliminaries

An FL system typically involves multiple distinct clients, denoted as $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$, and a central coordinating server C. Each S_k utilizes its own private dataset $D^k = \{(X^k, Y^k)\}$ to train a local model M_k. The server C aggregates the local model parameters $\{\theta_k\}$ from all participating clients to compute the global parameters $\theta_g = \sum_{k=1}^{K} \alpha_k \theta_k$, where α_k is the weight based on the size of each client's local dataset, and $\alpha_k = \frac{|D^k|}{\sum_{k=1}^{K} |D^k|}$.

Under this setting, FedDDL constructs a structural causal graph to analyze the confounding factors involved in FL model inference. As shown in Figure 2, the main idea is to tackle the interference of client data heterogeneity ($S \to X$)

S: Source, B: Background Feature, O: Object Feature, X: Image Feature, Y: Prediction
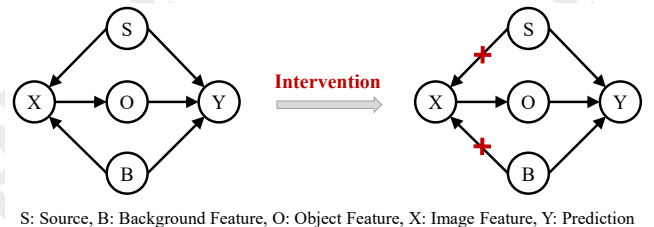
Figure 2: A causal view in federated out-of-distribution generalization, which uses backdoor intervention to eliminate the interference of background factor B and data source factor S on the sample X during model inference (i.e., $B \nrightarrow X$ and $S \nrightarrow X$).
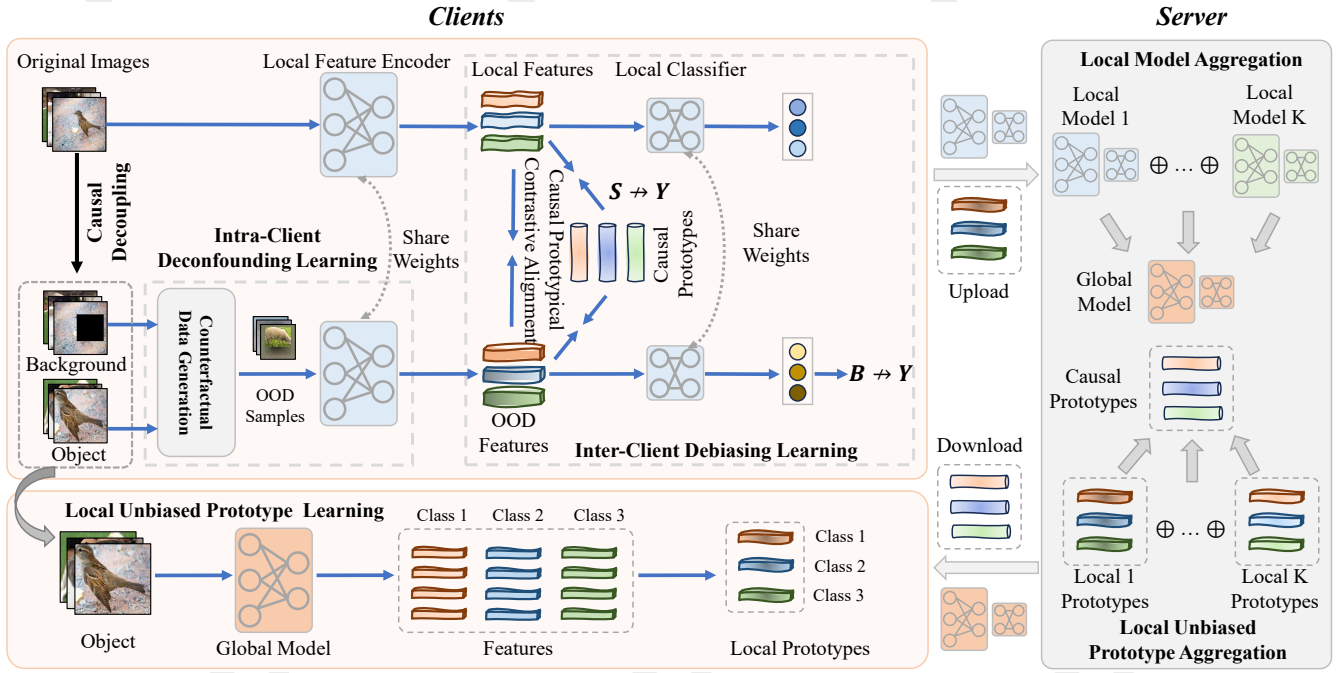
Figure 3: The `FedDDL` framework. It contains two main modules: 1) the intra-client deconfounding learning module, and 2) the inter-client debiasing learning module. The former generates counterfactual samples to break the spurious association between the background and specific labels $B \nrightarrow X$. The latter leverages causal prototypes to promote consistency in learning heterogeneous representations $S \nrightarrow X$.

and background ($B \rightarrow X$) factors in model decision-making, where X is the input image. The intra-client deconfounding learning module to generate counterfactual data to improve data diversity, which can sever the link between background and specific class during inference (i.e., $B \nrightarrow X$). The inter-client debiasing module constructs class-wise causal prototypes $U_c = u_c^1, \ldots, u_c^N$ from object images, and uses them to align heterogeneous representations across clients. This mitigates the influence of client-specific model differences during aggregation (i.e., $S \nrightarrow X$).

4 The Proposed `FedDDL` Method

This section presents the proposed `FedDDL` method with reference to the schematic framework shown in Figure 3.

4.1 A Causal View on FL OOD Generalization

To clarify the factors that influence model inference in federated OOD generalization, we have constructed a structured causal graph that comprehensively illustrates the potential inference paths of the model, as shown in Figure 2. Specifically,

- $X \leftarrow B \rightarrow Y$ indicates that the image background B is a potential confounding factor, causing the model to rely on it when inferring Label Y for Sample X. The proposed intra-client deconfounding learning module is designed to sever the connection between x and B (i.e., $X \nleftarrow B$).

- $X \leftarrow S \rightarrow Y$ implies that different clients contain diverse image features. Bridging this heterogeneity can eliminate interference between images and their labels. The proposed inter-client debiasing learning module is designed to

align heterogeneous representations, which severs the association between S and input image X (i.e., $X \nleftarrow S$).

- $X \rightarrow O \rightarrow Y$ means that the design of all modules focuses on learning the true causal effects, which infers labels Y based on key objects O in an image X, while avoiding spurious correlations with confounding factors B and S. By severing both paths, the causal structure is clearly revealed during federated OOD generalization.

4.2 Intra-Client Deconfounding Learning (DEC)

To address shortcut learning caused by client-specific backgrounds, the DEC module aims to block the causal link from background to label during inference, encouraging the model to focus on true causal relationships. To achieve this, it performs backdoor intervention to adjust the sample distribution $do(X)$. By applying the law of total probability, the inference involves both the direct causal impact of X on Y and the correlation confounded by B:

$$P(Y \mid X) = \sum_B P(Y \mid X, B)P(B \mid X). \tag{1}$$

To ensure that each background factor contributes equally to label inference, the DEC module generates counterfactual samples to intervene in the distribution:

$$\begin{aligned} P(Y \mid do(X)) &= \sum_B P(Y \mid DEC(X, B)) \\ &= \sum_B P(Y \mid X, B)P(B), \end{aligned} \tag{2}$$

which removes the path from B to X in Figure 2, thereby facilitating the model to approximate the causal intervention $P(Y \mid do(X))$ instead of the spurious correlation $P(Y \mid X)$. Specifically, the DEC module begins by decoupling images

Figure 4: Example counterfactual samples with causal features.

into objects and backgrounds. Then, it modifies the background factors associated with the objects. Inspired by the pre-trained model Grounding DINO [Liu and et al., 2025], the approach leverages text prompts T to detect targets:

$$\theta = \begin{bmatrix} \theta_1 & \theta_2 \\ \theta_3 & \theta_4 \end{bmatrix} = DINO(X, T), \tag{3}$$

$$I_O = I \odot 1_{(x,y) \in [\theta_1, \theta_2] \times [\theta_3, \theta_4]}, I_B = I \odot 1_{(x,y) \notin [\theta_1, \theta_2] \times [\theta_3, \theta_4]}, \tag{4}$$

where I_O and I_B are the object and background images. T is the label name. θ represents the bounding box coordinates (θ_1, θ_3 for the top-left, and θ_2, θ_4 for the bottom-right). $\odot$ denotes the Hadamard product. $1_{(x,y) \notin [x_1, x_2] \times [y_1, y_2]}$ is an indicator function, which evaluates to 1 outside the bounding box and 0 inside it.

Subsequently, to generate counterfactual samples with causal features and diverse backgrounds for distribution intervention, the DEC module divides the backgrounds I_B^i belonging to class i into η groups, denoted as $\{g_i^j | j = 1, \ldots, \eta\}$:

$$\{g_i^j | j = 1, \ldots, \eta\} = Split(I_B^i, \eta), \tag{5}$$

where $Split(\cdot)$ is the random grouping function. The counterfactual data I_C can be created by fusing randomly grouped backgrounds g_i^j and an object I_O:

$$I_C = \left[\frac{1}{|g_i^j|} \sum_{I_B \in g_i^j} I_B \right] \oplus Trans(I_O), \tag{6}$$

where $Trans(\cdot)$ denotes a transformation function, which may include rotation, flipping and resizing, as shown in Figure 4. To guide the model to focus on the target objects in different cases, we optimize it using a joint classification loss:

$$\mathcal{L}_J = - \sum_{n=1}^N y(n) \log (M_k(I)(n)) + - \sum_{n=1}^C y_C(n) \log (M_k(I_C)(n)), \tag{7}$$

where N is the total number of classes. $M_k(\cdot)$ is the model of client k. $y(n)$ and $y_c(n)$ are the true labels of $n - th$ index for the original and counterfactual samples. $M_k(I)(n)$ and $M_k(I_C)(n)$ are corresponding predictions.

4.3 Inter-Client Debiasing Learning (DEB)

To address the issue of conflicts among clients leading to declines in the overall FL model performance, the DEB module aims to sever the path from FL clients to labels in order to arrive at consistent decisions for the same input. It performs backdoor adjustments to the representation learning of each client. The original multi-source inference is expressed as:

$$P(Y|X) = \sum_{i=1}^K \sum_B P(Y|X, B, S_i)P(B|X, S_i)P(S_i|X), \tag{8}$$

Algorithm 1 FEDDDL

1: Initialize the global model parameter θ^0
2: **for** $t = 1, \ldots, T$ **do**
3:     Sample subset $\mathcal{K}$ of clients with $|\mathcal{K}| = k$
4:     **for** each client $k \in \mathcal{K}$ in parallel **do**
5:       Initialize local model parameter $\theta_k^t = \theta^{t-1}$
6:       Counterfactual sample generation $D_{C,k}$ based on objects $I_{O,k}$ and backgrounds $I_{B,k}$ images
7:       **for** $e = 1, \ldots, E$ **do**
8:         Sample batches of data ζ_1, ζ_2 from local data D^k and counterfactual data $D_{C,k}$
9:         **if** $t = 1$ **then**
10:          $g_k = \nabla \mathcal{L}_J(\zeta_1, \zeta_2)$
11:         **else**
12:          $g_k = \nabla \mathcal{L}_J(\zeta_1, \zeta_2) + \lambda \nabla \mathcal{L}_{CR}(\zeta_1, \zeta_2, U_G)$
13:         **end if**
14:         Update $\theta_k^t \leftarrow \theta_k^t - \eta_l g_k$
15:       **end for**
16:       $U_{L,t}^{i,k} \leftarrow \frac{1}{|I_O^{i,k}|} \sum M_G(\theta^{t-1}, I_O^{i,k}), i = 1, \ldots, N$
17:     **end for**
18:     $\theta^t = 1/k \sum_{k \in \mathcal{K}} \theta_k^{t-1}$
19:     $U_{G,t}^i = 1/k \sum_{k \in \mathcal{K}} U_{L,t}^{i,k}, i = 1, \ldots, N$
20: **end for**

---

which includes both the direct causal effect of X on Y and the correlation confounded by B and S. The DEC module mitigates the interference from factor B by severing the connection between the data source S and the label Y:

$$P(Y|do(X)) = \sum_{i=1}^k \sum_B P(Y|DEC(X,B), DEB(S_i)) \\ = \sum_{i=1}^k \sum_B P(Y|X, B, S_i)P(B)P(S_i), \tag{9}$$

Specifically, it first constructs causal prototypes using the extracted object images I_O to reduce the proportion of background features in the prototype components as:

$$U_L^{i,k} = \frac{1}{|I_O^{i,k}|} \sum_{\hat{I} \in I_O^{i,k}} M_G(\hat{I}), \tag{10}$$

where $U_L^{i,k}$ and $I_O^{i,k}$ represent the local causal prototype of class i and the data set in client k, respectively. $M_G(\cdot)$ denotes the global model. Moreover, the global causal prototype $U_G^{i,k}$ of class i can be represented as:

$$U_G^i = \frac{1}{K} \sum_{k=1}^K U_L^{i,k}. \tag{11}$$

As shown in Figure 5, backgrounds might cause different class prototypes to exhibit similar distributions across FL clients. In contrast, causal prototypes effectively eliminate the interference of backgrounds, thereby enhancing their distinction. Therefore, the DEB module performs causal prototype-based regularization to guide the adjustment of representation distributions in each client:

$$\mathcal{L}_{CR} = - \log \frac{\exp(f \cdot U_G^+/\tau)}{\exp(f \cdot U_G^+/\tau) + \sum \exp(f \cdot U_G^-/\tau)}, \tag{12}$$

where f represents a local feature. U_G^+ and U_G^- denote the global causal prototypes of the same and different classes
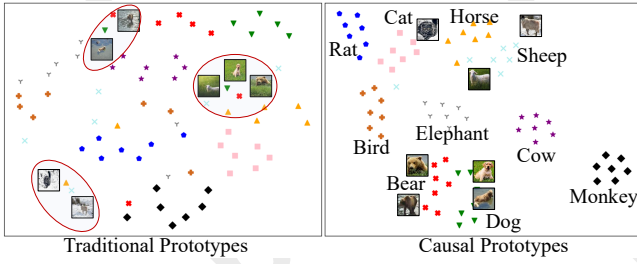
Figure 5: Comparison between traditional and causal prototypes. Causal prototypes alleviate the interference from specific attributes.

as f, respectively. τ is the temperature parameter. This reduces the output gap between models from different clients, and eliminates interference from clients' local factors.

4.4 Training Strategy

FedDDL aims to eliminate the interference of client-specific attributes and decision differences on the final model inference. This is achieved through the collaboration of the DEC module and the DEB module. We summarize the pseudo-code of FedDDL in Algorithm 1. Its overall optimization objective can be express as:

$$\mathcal{L}_{total} = E_{(x,y) \sim D_{local}}[\mathcal{L}_J + \lambda * \mathcal{L}_{CR}], \qquad (13)$$

where λ is a weighted parameter.

5 Experimental Evaluation

5.1 Experiment Settings

Datasets Following previous work on OOD generalization [Qi and et al., 2024; Wang *et al.*, 2022], experiments are conducted on two datasets: NICO-Animal and NICO-Vehicle [Wang *et al.*, 2021]. Their statistics and partitioning method of the datasets can be found in Table 1.

Datasets	#Class	#Training	#Testing
NICO-Animal (F7)	10	10,633	2,443
NICO-Animal (L7)	10	8,311	4,765
NICO-Vehicle (F7)	10	8,027	3,626
NICO-Vehicle (L7)	10	8,352	3,301

Table 1: Statistics of NICO-Animal and NICO-Vehicle. F7 represents data from the first seven backgrounds of each class used as the training set. L7 represents data from the last seven backgrounds of each class used as the training set.

Evaluation Metrics Following prior work [McMahan *et al.*, 2017; Li *et al.*, 2021], we use Top-1 Accuracy to evaluate performance, which is defined as $\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)$, where N is the total number of samples, $\hat{y}_i$ is the predicted label, and y_i is the ground-truth label.

Network Architecture To ensure fair comparison, all methods adopt the same network architecture. In FedDDL, a unified architecture is used to process both raw images and those processed with Grounding DINO. Following prior studies [**?**; Liu *et al.*, 2021], ResNet-18 [He *et al.*, 2016] is selected as the backbone for both datasets.

Implementation Details We set the local training epochs to 10 per global round for both datasets. The total number of communication rounds is 50, with 7 clients for both datasets. We used a client sampling fraction of 1.0 and employed SGD as the optimizer. During local training, the weight decay is set to 0.01, the batch size is 64, and the initial learning rate is 0.01 for both datasets. λ is chosen from the set $\{0.1, 1.0, 2.0\}$. τ is selected from $\{0.5, 0.07\}$. η is tuned from $\{1, 3, 5\}$. The BOX_THRESHOLD and TEXT_THRESHOLD in the DINO model have been set to 0.3. For other methods, hyperparameters are tuned based on the corresponding papers.

5.2 Performance Comparison

This section presents a comparison of the FedDDL method with nine state-of-the-art (SOTA) methods, including FedAvg [McMahan *et al.*, 2017], FedProx [Li *et al.*, 2020], MOON [Li *et al.*, 2021], FPL [Huang *et al.*, 2023], FedIIR [Guo *et al.*, 2023], FedDecorr [Shi *et al.*, 2024], FedHeal [Chen *et al.*, 2024], MCGDM [Wei and Han, 2024], and FedCCRL [Wang and Tang, 2024]. The results derived from Table 2 are summarized below.

- FedDDL achieve significant improvements across all cases compared to existing methods. This includes evaluations of both global (Global) and local (Local) models, where the consistent enhancements highlight the robustness of FedDDL in coping with complex scenarios.

- The DEC module is a plug-and-play component that can be easily integrated into other methods, such as MOON, and can provide significant performance improvements. This is reasonable because it enhances sample diversity and ensures the preservation of causal features.

- By addressing the federated OOD problem from both intra-client and inter-client perspectives, FedDDL outperforms single-perspective methods like MOON and FPL. FedDDL leverages deconfounding learning to disrupt spurious correlations between backgrounds and labels within clients, while also employing debiasing learning to reduce the output gap between clients.

- FedDDL also reduces the performance disparity between different local models (as reflected by the local variance). This can be understood as it simultaneously strengthens the performance of individual local models via the DEC module while leveraging the DEB module to promote consistency in outputs across heterogeneous clients.

5.3 Ablation Study

This section provides an in-depth analysis of the effectiveness of various modules within the FedDDL, including the intra-client deconfounding learning (DEC) module, the inter-client decbiasing learning module with local data (DEB_L), counterfactual data (DEB_C), and their combined set (DEB_{L+C}). The results are presented in Table 3.

- Incorporating the inter-client decbiasing learning module contributes to improving the performance of the global model and the average performance of the local models, but there is a significant variation in the performance across different client models.

Methods	NICO-Animal				NICO-Vehicle			
	Global		Local		Global		Local	
	F7	L7	F7	L7	F7	L7	F7	L7
FedAvg	44.38±0.6	52.75±0.6	34.19±4.6	44.45±4.6	63.28±0.4	59.05±0.2	48.44±6.9	44.79±6.2
Fedprox	44.55±0.9	51.99±0.9	35.72±4.9	43.17±5.3	65.36±0.6	57.50±0.8	47.62±5.1	42.63±5.3
MOON	45.53±0.4	53.66±0.9	36.31±5.7	46.12±6.2	65.94±0.5	59.63±0.5	51.34±5.8	46.54±5.2
FPL	47.76±0.5	55.39±0.2	37.58±4.5	46.93±4.4	68.51±0.7	61.76±0.6	52.22±5.1	48.26±4.9
FedDeccor	48.11±0.6	53.12±0.2	37.64±5.2	47.32±5.1	67.39±0.7	61.32±0.6	51.21±6.7	46.34±5.3
FedIIR	46.40±0.9	52.82±0.7	35.71±4.9	45.88±4.7	63.64±0.9	56.18±0.4	49.47±5.1	47.83±5.4
FedHeal	42.32±1.0	52.80±0.6	36.44±5.1	44.78±3.6	64.00±0.5	56.25±0.7	46.32±5.2	44.59±4.9
MCGDM	47.96±0.8	54.53±0.5	38.46±4.8	47.51±3.5	66.84±0.4	59.59±0.9	53.69±7.9	46.92±4.4
FedCCRL	48.49±0.7	57.31±0.9	39.81±3.9	47.37±4.2	70.14±0.5	62.23±0.9	54.73±6.8	50.32±7.6
MOON+DEC	52.47±0.9	60.13±0.7	42.56±5.9	50.26±6.1	71.43±0.6	64.45±1.1	56.32±4.4	52.56±5.7
FedDDL	**53.37±0.4**	**62.59±0.6**	**44.27±3.1**	**53.23±3.0**	**73.38±0.7**	**66.20±0.4**	**59.51±4.6**	**54.28±4.2**

Table 2: Performance comparison between FedDDL and baselines on NICO-Animal and NICO-Vehicle datasets. All methods were executed across three trials, and the results are reported as the mean and standard deviation of the top-1 accuracy.

	NICO-Animal (L7)		NICO-Vehicle (L7)	
	Global	Local	Global	Local
FedAvg	52.75±0.6	44.45±4.6	59.05±0.2	44.79±4.2
+ DEB_L	54.39±0.5	48.74±4.1	62.43±0.7	47.82±3.6
+ DEB_C	55.12±0.8	49.13±3.2	62.58±0.4	49.37±3.1
+ DEB_{L+C}	57.43±0.7	49.56±3.4	63.49±0.6	50.74±3.3
+ DEC	57.78±0.6	48.63±2.4	63.11±0.5	51.36±2.7
+ DEB_L + DEC	60.84±0.4	51.68±2.1	64.15±0.3	52.37±2.4
+ DEB_C + DEC	61.69±0.7	52.31±2.7	65.28±0.4	53.24±2.7
+ DEB_{L+C} + DEC	**62.54±0.5**	**53.23±2.7**	**66.20±0.4**	**54.28±2.2**

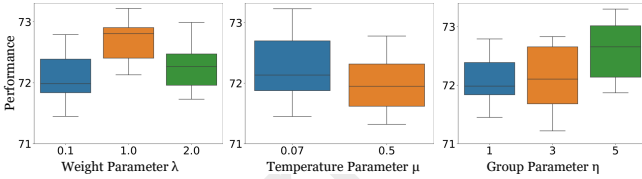Table 3: Ablation study of FedDDL on NICO-Animal (L7) and NICO-Vehicle (L7) with top-1 accuracy.

Figure 6: The impact of hyperparameters on performance. The λ, τ and η are tuned from $\{0.1, 1.0, 2.0\}$, $\{0.07, 0.5\}$ and $\{1, 3, 5\}$ on NICO_Vehicle (F7), respectively.

- The inter-client decbiasing learning module plays a significant role in bridging the performance gap between client models, which promotes consistent improvements across heterogeneous sources and enhances the alignment of model representations, diminishing discrepancies in client-specific performance.

- As the diversity of the data increases, the advantages of the DEB module are amplified, as it effectively enhances heterogeneous representation alignment, particularly in complex cases, where its ability to adapt and align diverse representations proves to be a significant asset.

5.4 Robustness of FedDDL on Hyperparameters

This section assesses the robustness of FedDDL under different hyperparameters. Specifically, we explore the impact of hyperparameters λ, τ, and η by selecting values from the sets $\{0.1, 1.0, 2.0\}$, $\{0.07, 0.5\}$, and $\{1, 3, 5\}$, respectively. As depicted in Figure 6, FedDDL consistently outperforms

FedAvg across various scenarios, showing remarkable insensitivity to hyperparameter variations within a broad range. This suggests that FedDDL is highly robust and adaptable to changes in hyperparameters. For the adjustment of τ, setting $\tau = 0.07$ enhances model performance compared to $\tau = 0.5$ by increasing the representation differences between sources. This adjustment allows the model to focus more on these variations, leading to improved generalization across clients. For η, increasing η means greater data diversity. The greater the diversity in the data, the more the model benefits, highlighting the importance of diverse data sources in improving the model's ability to generalize.

5.5 Case Study
Breaking the Background-Category Association
Here, we examine the effectiveness of the DEC module in breaking the association between background and specific labels. As shown in Figure 7(a), the unique attributes of the data lead to spurious associations in the FedAvg method between "dog" and "grass" as well as "cow" and "river". Even when the objects in the images are masked, the model still identifies the background with an exceptionally high confidence as the label of the object. This can easily cause the model to rely on non-causal associations to infer the labels of samples, ultimately leading to errors. Moreover, benefiting from the intervention of the DEC module, the association between background and labels is disrupted, leading the model to predict the background with equal probability across all dimensions. As shown in Figure 7(b), it narrows the gap among the probabilities of predicting the background as different categories, avoiding to use a very robust but causally wrong feature to make predictions.

Comparison of Visual Attention
This section analyzes the effectiveness of the deconfounding learning in local 1. We randomly selected test samples and visualized the visual attention of different methods using GradCAM [Selvaraju et al., 2017; Ślazyk et al., 2022; Lin et al., 2020; Meng et al., 2019]. As illustrated in Figure 8, FedAvg often focuses on the background of out-of-distribution samples, leading to classification errors due to its limited generalization ability. For example, in the case of a
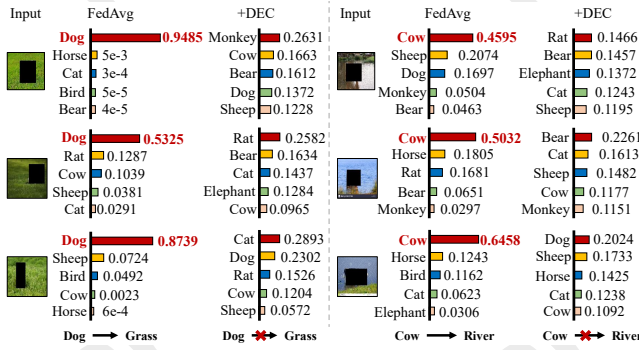
Figure 7: (a) FedAvg forms spurious associations between "dog" and "grass", as well as "cow" and "river"; (b) The DEC module helps the model reduce the probability gap between the background being predicted as any category.
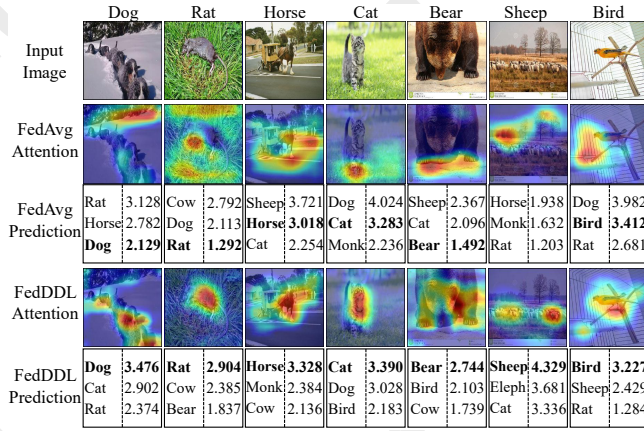

Figure 8: Visualization of visual attention and comparison of predictions for FedAvg and FedDDL. FedDDL enhances the model's ability to focus on target objects in out-of-distribution samples and improves the confidence in ground-truth predictions.

cat on grass, FedAvg may assign the label based on "grass," leading to incorrect identification as a "dog," since the training data only contains dogs on grass. Furthermore, FedDDL effectively focuses on the target object in test samples, enabling the model to eliminate background interference and make accurate predictions. This significantly enhances the generalization ability of local models and demonstrates that improving the performance of individual models also contributes to better collaboration.

Analysis of Cross-Silo Representation Alignment

This section analyzes the consistency of representation distributions from models of different sources on the same test samples. As shown in Figure 9, the image representations from clients 1, 2 and 3 are visualized using t-SNE [Van der Maaten and Hinton, 2008; Zhou *et al.*, 2023]. Clearly, the DEB module has mitigated the gap in outputs from models of different sources. Meanwhile, it has also enhanced the discriminability of representations across classes. This is reasonable because causal prototype contrastive alignment helps alleviate overfitting to specific attributes and facilitates the
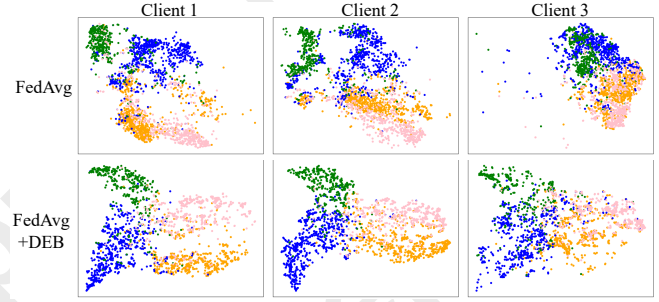

Figure 9: Visualization of representation distributions from different sources. The DEB module has made significant contributions to cross-silo feature alignment in out-of-distribution cases.

construction of a unified knowledge base across silos. In contrast, FedAvg may learn inconsistent inter-class relationships between different sources (e.g., the orange and pink classes in clients 1 and 2). In client 3, the coverage of different class representations is high, and the classification boundaries are not clearly defined. These observations demonstrate the effectiveness of the DEB module.

6 Conclusions and Future Work

In this paper, we address the problem of overfitting to specific backgrounds and the ill-posed aggregation issues caused by attribute skew in federated learning. We propose FedDDL, which comprehensively analyzes the factors interfering FL model inference. The key idea is to generate counterfactual samples to mitigate the interference of image backgrounds. Moreover, it promotes the consistency of outputs across models from different FL client through debiasing learning. Experimental results demonstrate that FedDDL significantly improves the performance of the global FL model in OOD cases.

In subsequent research, we plan to extend FedDDL to cases where attributes cannot be directly disentangled, and develop adaptive methods for causal relationship discovery.

Acknowledgments

References

[Cai *et al.*, 2024a] Jinyu Cai, Yunhe Zhang, and et al. Lg-fgad: An effective federated graph anomaly detection framework. In *IJCAI*, pages 3760–3769, 2024.

[Cai *et al.*, 2024b] Jinyu Cai, Yunhe Zhang, Zhoumin Lu, and et al. Towards effective federated graph anomaly de-

tection via self-boosted knowledge distillation. In *MM*, pages 5537–5546, 2024.

[Chen *et al.*, 2023] Junming Chen, Meirui Jiang, and et al. Federated domain generalization for image recognition via cross-client style transfer. In *WACV*, pages 361–370, 2023.

[Chen *et al.*, 2024] Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *CVPR*, pages 12077–12086, 2024.

[de Luca and et al., 2022] Artur Back de Luca and et al. Mitigating data heterogeneity in federated learning with data augmentation. *arXiv preprint arXiv:2206.09979*, 2022.

[Fan *et al.*, 2025] Tao Fan, Hanlin Gu, et al. Ten challenging problems in federated foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Fu *et al.*, 2025a] Lele Fu, Sheng Huang, and et al. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *AAAI*, pages 16648–16656, 2025.

[Fu *et al.*, 2025b] Lele Fu, Sheng Huang, Yanyi Lai, Chuanfu Zhang, and et al. Federated domain-independent prototype learning with alignments of representation and parameter spaces for feature shift. *IEEE Transactions on Mobile Computing*, pages 1–16, 2025.

[Guo *et al.*, 2023] Yaming Guo, Kai Guo, Xiaofeng Cao, and et al. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *ICML*, pages 11905–11933. PMLR, 2023.

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

[Hu *et al.*, 2023] Ming Hu, Zeke Xia, Dengke Yan, and et al. Gitfl: Uncertainty-aware real-time asynchronous federated learning using version control. In *RTSS*, pages 145–157. IEEE, 2023.

[Hu *et al.*, 2024] Ming Hu, Zeke Xia, Dengke Yan, and et al. Fedmut: Generalized federated learning via stochastic mutation. In *AAAI*, volume 38, pages 12528–12537, 2024.

[Huang *et al.*, 2023] Wenke Huang, Mang Ye, and et al. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322. IEEE, 2023.

[Huang *et al.*, 2025] Sheng Huang, Lele Fu, Yuecheng Li, Chuan Chen, and et al. A cross-client coordinator in federated learning framework for conquering heterogeneity. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8828–8842, 2025.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, and et al. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2021] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021.

[Liao *et al.*, 2024] Tianchi Liao, Lele Fu, Jialong Chen, and et al. A swiss army knife for heterogeneous federated learning: Flexible coupling via trace norm. *NeurIPS*, 37:139886–139911, 2024.

[Lin *et al.*, 2020] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. In *AAAI*, volume 34, pages 2661–2668, 2020.

[Liu and et al., 2025] Shilong Liu and et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55. Springer, 2025.

[Liu *et al.*, 2021] Quande Liu, Cheng Chen, Jing Qin, and et al. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021.

[Liu *et al.*, 2024] Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. Fedlpa: One-shot federated learning with layer-wise posterior aggregation. *Advances in Neural Information Processing Systems*, 37:81510–81548, 2024.

[Luo *et al.*, 2022] Zhengquan Luo, Yunlong Wang, Zilei Wang, and et al. Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring. *arXiv preprint arXiv:2206.06818*, 2022.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, and et al. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282. PMLR, 2017.

[Meng *et al.*, 2019] Lei Meng, Long Chen, and et al. Learning using privileged information for food recognition. In *MM*, pages 557–565, 2019.

[Meng *et al.*, 2024] Lei Meng, Zhuang Qi, Lei Wu, Du Xiaoyu, and et al. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36, 2024.

[Morafah and et al., 2024] Mahdi Morafah and et al. Stable diffusion-based data augmentation for federated learning with non-iid data. *arXiv preprint arXiv:2405.07925*, 2024.

[Mu *et al.*, 2023] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, and et al. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143:93–104, 2023.

[Nguyen *et al.*, 2024] Dung Thuy Nguyen, Taylor T Johnson, and Kevin Leach. Fisc: Federated domain generalization via interpolative style transfer and contrastive learning. *arXiv preprint arXiv:2410.22622*, 2024.

[Park *et al.*, 2024] Jungwuk Park, Dong-Jun Han, and et al. Stablefdg: style and attention based learning for federated domain generalization. *NeurIPS*, 36, 2024.

[Qi and et al., 2024] Zhuang Qi and et al. Attentive modeling and distillation for out-of-distribution generalization of federated learning. In *ICME*, pages 1–6. IEEE, 2024.

[Qi *et al.*, 2022] Zhuang Qi, Yuqing Wang, Zitan Chen, Ran Wang, Xiangxu Meng, and Lei Meng. Clustering-based curriculum construction for sample-balanced federated learning. In *CICAI*, pages 155–166. Springer, 2022.

[Qi *et al.*, 2023] Zhuang Qi, Lei Meng, Zitan Chen, and et al. Cross-silo prototypical calibration for federated learning with non-iid data. In *MM*, pages 3099–3107, 2023.

[Qi *et al.*, 2024] Zhuang Qi, Lei Meng, Weihao He, Ruohan Zhang, and et al. Cross-training with multi-view knowledge fusion for heterogenous federated learning. *arXiv preprint arXiv:2405.20046*, pages 1–12, 2024.

[Qi *et al.*, 2025a] Zhuang Qi, Lei Meng, and et al. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *AAAI*, pages 19986–19994, 2025.

[Qi *et al.*, 2025b] Zhuang Qi, Runhui Zhang, and et al. Global intervention and distillation for federated out-of-distribution generalization. In *ICME*, pages 1–6, 2025.

[Ren *et al.*, 2025] Chao Ren, Han Yu, Hongyi Peng, Xiaoli Tang, Bo Zhao, Liping Yi, et al. Advances and open challenges in federated foundation models. *IEEE Communications Surveys and Tutorials*, 2025.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, and et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[Shenaj *et al.*, 2023] Donald Shenaj, Eros Fanì, Marco Toldo, and et al. Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning. In *WACV*, pages 444–454, 2023.

[Shi *et al.*, 2024] Yujun Shi, Jian Liang, Wenqing Zhang, Chuhui Xue, and et al. Understanding and mitigating dimensional collapse in federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2936–2949, 2024.

[Ślazyk *et al.*, 2022] Filip Ślazyk, Przemysław Jabłecki, Aneta Lisowska, Maciej Malawski, and et al. Cxr-fl: deep learning-based chest x-ray image analysis using federated learning. In *ICCS*, pages 433–440. Springer, 2022.

[Sun *et al.*, 2023] Yuwei Sun, Ng Chong, and Hideya Ochiai. Feature distribution matching for federated domain generalization. In *ACML*, pages 942–957. PMLR, 2023.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J MACH LEARN RES*, 9(11), 2008.

[Wang and Tang, 2024] Xinpeng Wang and Xiaoying Tang. Fedccrl: Federated domain generalization with cross-client representation learning. *arXiv preprint arXiv:2410.11267*, 2024.

[Wang *et al.*, 2021] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *CVPR*, pages 3091–3100, 2021.

[Wang *et al.*, 2022] Yuqing Wang, Xiangxian Li, and et al. Meta-causal feature learning for out-of-distribution generalization. In *ECCV*, pages 530–545. Springer, 2022.

[Wang *et al.*, 2023] Haozhao Wang, Yichen Li, Wenchao Xu, and et al. Dafkd: Domain-aware federated knowledge distillation. In *CVPR*, pages 20412–20421, 2023.

[Wang *et al.*, 2024] Haozhao Wang, Haoran Xu, Yichen Li, and et al. Fedcda: Federated learning with cross-rounds divergence-aware aggregation. In *ICLR*, 2024.

[Wei and Han, 2024] Yikang Wei and Yahong Han. Multi-source collaborative gradient discrepancy minimization for federated domain generalization. In *AAAI*, volume 38, pages 15805–15813, 2024.

[Xu *et al.*, 2023] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yi-Yan Wu, and et al. Federated adversarial domain hallucination for privacy-preserving domain generalization. *IEEE Transactions on Multimedia*, 26:1–14, 2023.

[Yang *et al.*, 2020] Qiang Yang, Lixin Fan, and Han Yu. Federated learning: Privacy and incentive, 2020.

[Yi *et al.*, 2023] Liping Yi, Gang Wang, and et al. Fedgh: Heterogeneous federated learning with generalized global header. In *MM*, pages 8686–8696. ACM, 2023.

[Yi *et al.*, 2024] Liping Yi, Han Yu, Chao Ren, et al. Federated model heterogeneous matryoshka representation learning. In *NeurIPS*, 2024.

[Yu and et al., 2011] Han Yu and et al. Dynamic witness selection for trustworthy distributed cooperative sensing in cognitive radio networks. In *ICCT*, pages 1–6, 2011.

[Yu *et al.*, 2023] Xinhui Yu, Dan Wang, Martin J. McKeown, and et al. Contrastive-enhanced domain generalization with federated learning. *IEEE Transactions on Artificial Intelligence*, 5(4):1525–1532, 2023.

[Zhang and et al., 2024] Jiayuan Zhang and et al. Enabling collaborative test-time adaptation in dynamic environment via federated learning. In *KDD*, pages 4191–4202, 2024.

[Zhang *et al.*, 2024] Xunzheng Zhang, Juan Marcelo Parra-Ullauri, Shadi Moazzeni, and et al. Federated analytics with data augmentation in domain generalization towards future networks. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[Zhang *et al.*, 2025a] Kaiyan Zhang, Jiayuan Zhang, Haoxin Li, and et al. Openprm: Building open-domain process-based reward models with preference trees. In *ICLR*, 2025.

[Zhang *et al.*, 2025b] Runhui Zhang, Sijin Zhou, and Zhuang Qi. Federated out-of-distribution generalization: A causal augmentation view. *arXiv preprint arXiv:2504.19882*, 2025.

[Zhao *et al.*, 2023] Zhuang Zhao, Feng Yang, and et al. Federated learning based on diffusion model to cope with non-iid data. In *PRCV*, pages 220–231. Springer, 2023.

[Zhou *et al.*, 2023] Tailin Zhou, Jun Zhang, and Danny HK Tsang. Fedfa: Federated learning with feature anchors to align features and classifiers for heterogeneous data. *IEEE Transactions on Mobile Computing*, 2023.

[Zhu *et al.*, 2024] Guogang Zhu, Xuefeng Liu, Jianwei Niu, and et al. Dualfed: enjoying both generalization and personalization in federated learning via hierachical representations. In *MM*, pages 11060–11069, 2024.