FedSDWC: Federated Synergistic Dual-Representation Weak Causal Learning for OOD

Zhenyuan Huang [1] *, Hui Zhang [1] *, Wenzhong Tang [1] *, Haijun Yang [1] *†

[1] Beihang University

zhenyuan@buaa.edu.cn, hzhang@buaa.edu.cn, tangwenzhong@buaa.edu.cn, navy@buaa.edu.cn

Abstract

Amid growing demands for data privacy and advances in computational infrastructure, federated learning (FL) has emerged as a prominent distributed learning paradigm. Nevertheless, differences in data distribution (such as covariate and semantic shifts) severely affect its reliability in real-world deployments. To address this issue, we propose FedSDWC, a causal inference method that integrates both invariant and variant features. FedSDWC infers causal semantic representations by modeling the weak causal influence between invariant and variant features, effectively overcoming the limitations of existing invariant learning methods in accurately capturing invariant features and directly constructing causal representations. This approach significantly enhances FL's ability to generalize and detect OOD data. Theoretically, we derive FedSDWC's generalization error bound under specific conditions and, for the first time, establish its relationship with client prior distributions. Moreover, extensive experiments conducted on multiple benchmark datasets validate the superior performance of FedSDWC in handling covariate and semantic shifts. For example, FedSDWC outperforms FedICON, the next best baseline, by an average of 3.04% on CIFAR-10 and 8.11% on CIFAR-100.

Introduction

Federated Learning (FL) has emerged as a key distributed learning paradigm for its ability to enable collaborative model training while preserving data privacy (McMahan et al. 2017; Liao et al. 2024; Zheng et al. 2020; Kumar et al. 2025). However, its practical application is hindered by significant challenges, most notably the non-independent and identically distributed (non-IID) nature of client data (Li et al. 2020). Beyond data heterogeneity, a more pressing challenge is out-of-distribution (OOD) generalization (Jiang and Lin 2022; Sefidgaran et al. 2024; Qi et al. 2025). In real-world FL, training occurs on a subset of clients and their data, creating a distributional shift—known as covariate shift—between the training data and the true data population (as shown in Fig. 1(b)). This phenomenon is known as the OOD generalization problem in FL (Liao et al. 2024; Zhou et al. 2025), where the core task is to capture stable feature-label relationships under covariate shift, enabling the model to generalize effectively to unseen clients. Under this context, FL must also address the semantic shift problem, which involves identifying data samples that do not belong to known categories during training. For example, in Fig. 1(c), categories such as cows and horses have not been encountered during training. Therefore, the model should be capable of rejecting such data rather than misclassifying them into known categories. Concurrently, FL models must handle semantic shifts, where they need to perform OOD detection to identify and reject data from new categories not seen during training (e.g., identifying a cow when trained only on cats and dogs in Fig. 1(c)), rather than misclassifying them.

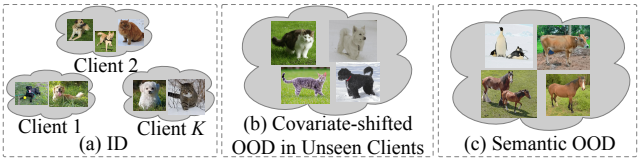


Figure 1: FL faces three primary data challenges, illustrated by a cat and dog identification task: (1) In-Distribution (ID) data is the heterogeneous training data from participating clients. This data often has non-identical distributions, such as one client having images of dogs on grass while another has cats in the snow. (2) Covariate-Shifted data is new data where the features of existing classes have changed. For example, the background for a dog shifts from grass to snow, while a cat's background shifts from snow to grass. (3) Semantic-Shifted data contains new categories, such as cows and horses, that were not in the training set.

Existing research on OOD generalization, including disentangled learning (Lee, Yao, and Finn 2023; Wang et al. 2023; Bai et al. 2024), invariant risk minimization (IRM) (Arjovsky et al. 2019; Li et al. 2022; Guo et al. 2023), and causal inference (Liu et al. 2021; Bravo-Hermsdorff et al. 2024; Noohdani et al. 2024; Zhang, Zhou, and Qi 2025) primarily aims to extract invariant features that are stable across different data distributions. However, these methods are often limited by the restrictive assumption that a representation function $\Psi(\cdot)$ exists, which can extract features from the in-

*These authors contributed equally.

†Corresponding Author.

put data X that are perfectly invariant across any two environments (e, e'), such that $\Psi(X^e) = \Psi(X^{e'})$. This assumption is difficult to satisfy in practice. Furthermore, by focusing exclusively on invariant features, these approaches risk discarding valuable information contained within the variant features of the data. Directly extracting causal features from complex data like images also remains a significant hurdle.

To address these limitations, we relax the strict invariance assumption, instead assuming that our feature mapping can extract most, but not all, invariant information. Crucially, we incorporate the often-overlooked variant features, positing they contain a smaller but still valuable amount of information. Based on this, we construct a causal graph (Fig. 1) between the latent factors derived from both invariant and variant features, bypassing the difficulty of direct causal discovery from raw data.

We summarize our main contributions as follows:

1. We propose FedSDWC, a novel FL model grounded in weak causal inference. It simultaneously addresses OOD generalization and detection by fusing invariant and variant features through a modeled weak causal relationship. Crucially, our approach relaxes the strict invariance assumptions common in traditional methods, enabling more flexible and robust feature utilization.
2. We design a novel intervention-based learning strategy to capture the weak causal dependency between invariant and variant features. Theoretically, we provide a tight generalization error bound for FedSDWC. More significantly, we are the first to establish a formal connection between the generalization of causal representations in FL and the clients' prior distributions, addressing a key gap in the literature.
3. Extensive experiments on multiple benchmark datasets demonstrate that FedSDWC significantly outperforms state-of-the-art (SOTA) models in both OOD generalization and detection, particularly under complex scenarios.

Related Works

Federated Learning with Non-IID Data

Data heterogeneity across clients is a primary obstacle in FL, significantly degrading the performance of standard algorithms like FedAvg (McMahan et al. 2017). One line of work aims to create a more robust global model. For instance, FedProx (Li et al. 2020) introduces a regularization term to constrain client model deviation, while FRAug (Chen et al. 2023) uses representation augmentation with a shared generator to capture cross-client consistency. A parallel approach is personalized FL, which tailors models to individual clients. Methods like FedL2P (Lee et al. 2024) leverage a meta-network to learn client-specific parameters, and pFedBreD (Shi et al. 2024) decouples personalized priors to improve adaptability. While effective for data heterogeneity, these methods lack dedicated mechanisms for OOD data, limiting their generalization performance.

OOD Generalization and Detection

OOD generalization aims to learn robust models by extracting invariant feature-label relationships amidst covariate shifts to ensure reliable deployment (Lv et al. 2023; Liao et al. 2024; Nguyen et al. 2025). Research in this area primarily follows three paradigms. Invariant learning, such as IRM, seeks invariant representations across different training environments (Arjovsky et al. 2019; Guo et al. 2023; Li et al. 2022). Disentangled learning separates data into stable and variant semantic factors to build robust representations (Kong et al. 2022; Bai et al. 2024), while causal inference leverages tools like structural causal models to eliminate spurious correlations (Gui et al. 2024; Zhang et al. 2025; Guo et al. 2025). Complementary to this, OOD detection focuses on identifying unknown or novel samples during inference. Prominent approaches include classification-based methods, which utilize model outputs like softmax probabilities (Djurisic et al. 2023; Linderman et al. 2023; Park, Jung, and Teoh 2023; Hendrycks and Gimpel 2016); distance-based methods, which measure a sample's distance to class prototypes, often using the Mahalanobis distance (Sun et al. 2022; Galesso, Argus, and Brox 2023; Ming and Li 2024); and density-based methods, which model the ID data with techniques like variational autoencoders or flows (Wang et al. 2022; Yang, Zhou, and Liu 2023; Wu and Deng 2023). Existing methods treat OOD generalization and detection as separate tasks and are constrained by strict invariance assumptions. We overcom these limitations with a unified, weak causal framework that leverages both invariant and variant features to improve performance on both tasks simultaneously.

Methodology

Problem Setting

OOD Generalization and Detection Objective in FL. In real-world FL deployment scenarios, each client c possesses its own dataset $\mathcal{D}_c$, leading to two possible cases: 1) The client c participating in training may have limited data available for model training due to various reasons, such as a large volume of data or the addition of new clients later. The data used for training is referred to as $\mathcal{D}_c^{\text{ID}}$. The remaining data may contain covariate-shifted data $\mathcal{D}_c^{\text{ID-C}}$ and semantic-shifted data $\mathcal{D}_c^{\text{ID-S}}$. 2) For clients not participating in training, their data composition is similar to that of participating clients and may also contain the three aforementioned types of components. Thus, the dataset of client c can be represented as $\mathcal{D}_c = \mathcal{D}_c^{\text{ID}} + \mathcal{D}_c^{\text{ID-C}} + \mathcal{D}_c^{\text{ID-S}}$. Our objective is:

$$\arg\min_{\theta} \sum_{c=1}^{C} w_c \mathbb{E}_{x \sim p_{\mathcal{D}_c}} \left[\mathcal{L}_c(\theta; \mathcal{D}_c) \right], \tag{1}$$

where w_c denotes the weight proportion of the c-th client, and θ represents the model parameters, including those of the classification model and the detector. The $\mathcal{L}_c(\theta; \mathcal{D}_c)$ can be further decomposed into three components: $\mathcal{L}_c(\theta; \mathcal{D}_c) = \ell_c^{\text{ID}} + \ell_c^{\text{ID-C}} + \ell_c^{\text{ID-S}}$, where ℓ_c^{ID} evaluates the generalization performance of client c on the $\mathcal{D}_c^{\text{ID}}$, $\ell_c^{\text{ID-C}}$ evaluates the generalization performance on covariate-shifted data $\mathcal{D}_c^{\text{ID-C}}$, and $\ell_c^{\text{ID-S}}$ evaluates the detection performance on semantic-shifted data $\mathcal{D}_c^{\text{ID-S}}$. The specific definitions are as follows:

$$\ell_c^{\text{ID}} := -\mathbb{E}_{(x,y) \sim p_{\mathcal{D}_c^{\text{ID}}}} \left[\mathbb{I}\{ y_{\text{pred}}(f_\theta(x)) = y \} \right],$$

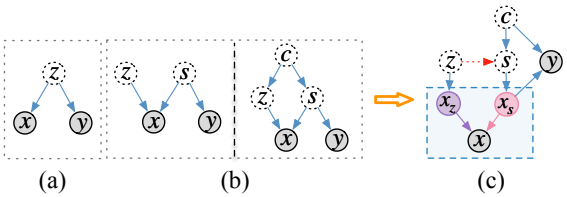


Figure 2: Comparison of different causal structures: (a) The observed variables x and y are solely influenced by the latent variable z, forming a simple causal structure. (b) Based on the original structure, an additional latent variable s is introduced, assuming that y is solely influenced by s, or assuming that both z and s are influenced by a deeper latent variable c. (c) FedSDWC improves on the existing models by decomposing x into invariant features x_s and environment-related variant features x_z, which are controlled by s and z, respectively. In inferring s, we consider both the invariant features of x_s and the weak causal influence of z. Finally, the c is derived through s, and x_s is used to infer $p(y|c, x_s)$.

$$\ell_c^{\text{ID-C}} := -\mathbb{E}_{(x,y)\sim p_{\mathcal{D}_c^{\text{ID-C}}}}\left[\mathbb{I}\{y_{\text{pred}}(f_\theta(x)) = y\}\right],$$
$$\ell_c^{\text{ID-S}} := -\mathbb{E}_{x\sim p_{\mathcal{D}_c^{\text{ID-S}}}}\left[\mathbb{I}\{g_\theta(x) = \text{ID}\}\right],$$

where $\mathbb{I}(\cdot)$ is the indicator function, $f_\theta(\cdot)$ is the classification model, and $g_\theta(\cdot)$ is the detector.

Design of Causal Model in Federated Learning

Our model design is grounded in a formal definition of causality: *a causal relationship exists between two variables (denoted as "cause → effect") if and only if intervening on the cause (by altering external variables of the system) can potentially change the effect, while the reverse does not hold* (Peters, Janzing, and Schölkopf 2017; Liu et al. 2021). Based on this definition of causality, we build our method by step-by-step refining a baseline model, using the image (x) to label (y) generation process as an example (Fig. 2(c)).

(1) **Initial Latent Structure.** In traditional discriminative models, the relationship $x \to y$ is typically learned directly. However, causal inference focuses more on the latent factors hidden between x and y. Therefore, we assume that the association between x and y mainly originates from a latent variable z (i.e., a "pure common cause"). Based on this assumption, the causal graph removes the direct edge $x \to y$ and retains only the indirect path through z (Fig. 2(a)).

(2) **Decomposition of Latent Variables.** The initial latent variable z is further decomposed to separate its causal and non-causal components. It is split into a semantic factor s (e.g., shape), which is the direct cause of the label y, and a variant factor z (e.g., background), which captures diversity in the input x. This refinement is represented in the causal graph by removing the direct edge from $z \to y$, thereby isolating the true causal pathway more precisely (Fig. 2(b)).

(3) **Eliminating Spurious Correlations.** Although s and z may exhibit certain statistical correlations in the data (e.g., camels/horses often appear in desert/grassland backgrounds), these are usually spurious correlations. For example, placing a horse in a desert background does not change its label. Therefore, we explicitly distinguish between s and z in the causal graph to eliminate their spurious correlations. However, it is unrealistic to completely infer s and z from the raw data x, as x typically contains confounding noise.

To overcome the challenge of inferring latent factors directly from x, the framework first extracts intermediate feature representations. Crucially, unlike methods that discard variant information, this model utilizes both an invariant feature representation x_s and a variant feature representation x_z to serve as the basis for inferring s and z. Recognizing that this feature separation is inevitably imperfect, the model introduces its central innovation: a weak causal relationship ($z \dashrightarrow s$) (Fig. 2(c)). This edge provides a principled way to model and account for semantic information that may have leaked into the variant features, thereby enhancing the model's causal reasoning and generalization performance.

Method for OOD Generalization and Detection

During training, the model leverages only the ID data, $\mathcal{D}_c^{\text{ID}}$, from each participating client $c \in \mathcal{C}_{par}$. Through server-side aggregation, all clients fit the global causal graph $p := \langle p(s|z,c), p(x_s|s), p(x_z|z), p(x|x_s,x_z), p(y|x_s,c)p(c)p(z)\rangle$ by maximizing the likelihood $\sum_{c\in\mathcal{C}_{par}} \mathbb{E}_{p_c^*(x,y)}[\log p(x,y)]$. The design of this model is based on the well-known independent causal mechanisms principle. However, in the FL setting, we make certain adaptations to this principle.

Federated Learning Causal Invariance: The proposed causal generative mechanisms $p(x|c,z)$ and $p(y|x_s,c)$ remain invariant across all clients and domains, while domain shifts are reflected solely through $p(z)$. It is infeasible for each client c to directly maximize the likelihood $\mathbb{E}_{p_c^*(x,y)}[\log p(x,y)]$, since $p(x,y) := \int p(c,s,z,x_z,x_s,x,y)\, dcdsdzdx_zdx_s$, where $p(c,s,z,x_z,x_s,x,y) := p(c)p(z)p(s|c,z)p(x_s|s)p(x_z|z)p(x|x_z,x_s)p(y|x_s,c)$, which is difficult to estimate directly. To address this issue, the Evidence Lower Bound (ELBO), defined as $\mathcal{L}_{p,q_{x_s,x_z,z,s,c|x,y}}(x,y) := \mathbb{E}_{q(x_s,x_z,z,s,c|x,y)}\left[\log \frac{p(x_s,x_z,z,s,c|x,y)}{q(x_s,x_z,z,s,c|x,y)}\right]$, can be introduced as an alternative optimization objective. By introducing an inference model $q(x_s,x_z,z,s,c|x,y)$, the process of sampling and density estimation can be simplified. Maximizing the ELBO enables $q(x_s,x_z,z,s,c|x,y)$ to approximate the posterior distribution $p(x_s,x_z,z,s,c|x,y) := \frac{p(x_s,x_z,z,s,c)}{p(x,y)}$, providing a tighter lower bound for optimizing $\log p(x,y)$.

However, even after introducing the inference model $q(x_s,x_z,z,s,c|x,y)$, it remains challenging to directly estimate $p(y|x)$, making the prediction task difficult. To address this issue, we further introduce the model $q(x_s,x_z,z,s,c,y|x)$ to approximate the target distribution $p(x_s,x_z,z,s,c,y|x)$. Through this approximation, it becomes possible to estimate y by sampling given x. Specifically, we further transform $q(x_s,x_z,z,s,c|x,y)$ and express it as: $q(x_s,x_z,z,s,c|x,y) = \frac{q(x_s,x_z,z,s,c,y|x)}{q(y|x)}$, where $q(y|x) := \int q(x_s,x_z,z,s,c,y|x)\, dcdsdzdx_zdx_s$, which is entirely determined by $q(x_s,x_z,z,s,c,y|x)$. Thus, the

ELBO objective for each client c, $\mathbb{E}_{p_c^*(x)}[\mathcal{L}_p, q_{s,z|x,y}(x,y)]$, can be formulated as:

$$\mathbb{E}_{p_c^*(x)}\mathbb{E}_{p^*(y|x)} \log q(y|x) + \mathbb{E}_{p^*(x)}\mathbb{E}_{q(c,s,z,x_z,x_s,y|x)} \left[\frac{p^*(y|x)}{q(y|x)} \log \frac{p(c,s,z,x_z,x_s,x,y)}{q(c,s,z,x_z,x_s,y|x)}\right]. \tag{2}$$

The first term of the objective function is the negative of the cross-entropy loss, which drives $q(y|x)$ closer to $p^*(y|x)$. As this goal is gradually achieved, the second term becomes the ELBO expectation $\mathbb{E}_{p_c^*(x)}[\mathcal{L}_{p,q_{(x_s,x_z,z,s,c,y|x)}}(x)]$, which works to further approximate $q_{(x_s,x_z,z,s,c,y|x)}$ to $p(x_s,x_z,z,s,c,y|x)$, and ensures $p(x)$ approaches $p^*(x)$. Moreover, based on our causal graph (Fig. 2(c)), the target distribution can be further decomposed as: $p(x_s,x_z,z,s,c,y|x) = p(x_s|x)p(x_z|x)p(s|x_s)p(z|x_z,s)p(y|c,x_s)$, where $p(y|c,x_s)$ is a known part of the model. Therefore, we can simplify $q_{(x_s,x_z,z,s,c,y|x)}$ using the inference models $q(x_s|x)$, $q(x_z|x)$, $q(s|x_s)$, and $q(z|x_z,s)$. This allows us to further rewrite Equation (2) as:

$$\mathbb{E}_{p_c^*(x)}[\mathcal{L}_p, q_{s,z|x,y}(x,y)] = \mathbb{E}_{p^*(x,y)} \log q(y|x) + \mathbb{E}_{p^*(x,y)} \left[\frac{1}{q(y|x)} \mathbb{E}_{q(c,s,z,x_z,x_s|x)} p(y|c,x_s) \cdot \log \frac{p(s|z,c)p(z)p(c)p(x_z|z)p(x_s|s)p(x|x_z,x_s)}{q(c,s,z,x_z,x_s|x)}\right]. \tag{3}$$

The above expectations can be estimated using the reparameterization trick combined with the Monte Carlo method. The objective function for client c is expressed as $\mathcal{L}_{elbo}^c$. The detailed derivation can be found in Appendix A.

Interventional Learning of Weak Causality

In our model, the relationship between the latent factors—the variant-derived z and the invariant-derived s—is learned through ELBO-based optimization after they are disentangled from the input x. It is important to note that we relax a commonly used assumption in the invariant learning literature (Arjovsky et al. 2019; Guo et al. 2023), which states that there exists a representation function $\Phi(\cdot)$ such that for all clients $c, c' \in \mathcal{C}_{\text{all}}$ and for any $\hat{z}$ within the intersection of the support sets $\text{supp}(\mathbb{P}(\Phi(X^c)))\cap \text{supp}(\mathbb{P}(\Phi(X^{c'})))$, the following relationship holds:

$$\mathbb{E}_{X^c,Y^c}[Y^c|\Phi(X^c) = z] = \mathbb{E}_{X^{c'},Y^{c'}}[Y^{c'}|\Phi(X^{c'}) = z].$$

We argue that achieving such perfect feature disentanglement is impractical in real-world scenarios. Instead, we posit that the variant (x_z) and invariant (x_s) features can only be partially decoupled. This leads to our central hypothesis: a weak causal relationship exists from z to s, capturing the subtle but non-negligible influence that arises from this imperfect separation. The key challenge is to learn this weak causal relationship effectively. To address this, we design a novel, intervention-based objective function aimed at capturing the system's response to small perturbations of the variant factor. We formalize this as the interventional consistency loss, denoted as $\mathcal{L}_{ic}$. Specifically, for any given sample x, we first decouple it into its invariant feature x_s and variant feature x_z. The intervention involves perturbing the variant feature x_z with scaled gaussian noise, yielding the intervened feature $\hat{x}_z = x_z + \alpha\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The loss penalizes the divergence between predictions on (x_s, x_z) and $(x_s, \hat{x}_z)$. We define the loss using KL divergence to measure the discrepancy:

$$\mathcal{L}_{ic} = \mathbb{E}_{x\sim\mathcal{D}}\left[\text{KL}\left(q(y|x_s,x_z) \,||\, q(y|x_s,\hat{x}_z)\right)\right] \tag{4}$$

By minimizing $\mathcal{L}_{ic}$, we compel the model to produce consistent predictions that are robust to α-scaled perturbations in the variant features. This directly constrains the influence flowing through the $z \dashrightarrow s$ causal pathway. The scaling factor α acts as a crucial hyperparameter, controlling the strength of this regularization: a larger α enforces a stricter invariance, pushing the model to learn a weaker causal link.

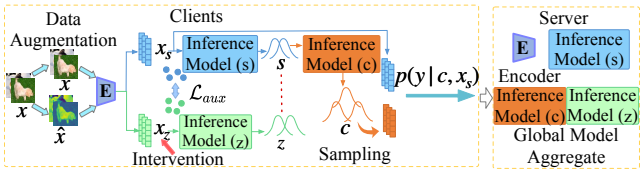


Figure 3: Framework of FedSDWC.

Model Framework

The FedSDWC model architecture, depicted in Fig. 3. On the client side, input data x first undergoes Fourier augmentation (Xu et al. 2021). This augmented data is then fed into an encoder (WideResNet for CIFAR datasets, ResNet-18 for TinyImageNet), in conjunction with an auxiliary learning method. During training, an intervention strategy is applied to improve generalization. Three MLP-based inference models estimate the distributions of latent factors (s, z, and c) using Gaussian Mixture Models. Finally, the server aggregates the updated client models using the FedAvg algorithm. The complete training procedure is detailed in Algorithm 1.

Theoretical Analysis

Assumption 1 (Additive Noise). According to causal graph (Fig. 2), the data x and y for each client follow the conditional distributions $p(x|c,z) = p_\mu(x - f(c,z))$ and $p(y|c,z) = p_\epsilon(y - h(c,z))$, where μ and ϵ are independent random variables. The function f is bijective, while h is injective. For categorical variables y, the conditional distribution can be expressed as $p(y|c,z) = \text{Cat}(y \mid h(c,z))$. Furthermore, the nonlinear functions f and h have bounded third-order derivatives.

Assumption 2. The noise distribution p_μ for each client has an almost everywhere (a.e.) nonzero characteristic function, such as a Gaussian distribution.

Theorem 1 (OOD Generalization Error). Under the conditions of Assumptions 1 and 2, the globally aggregated causal model p_g in FL and the OOD model $\tilde{p}$ share the same generative mechanism. However, due to environmental differences, the prior distributions p_k of different clients may vary. For all client data $x \in \text{supp}(p_{k,x}) \cap \text{supp}(\tilde{p}_x)$, where

Algorithm 1: Training procedure of FedSDWC

Input: Communication rounds T, set of participating clients $\mathcal{C}_{\text{par}}$, local steps E, batch size B.
Ouput: Optimized global model parameters θ_T.
Server Executes.
Initialize global model parameters (θ_T).
for $t = 1$ **to** T **do**
  for $c \in \mathcal{C}_{\text{par}}$ in parallel do **do**
    Send the model θ_t to the client c.
    $\theta_t^c \leftarrow$ Client Executes(c, θ_t).
  end for
  $\theta_{t+1} = \sum_{c \in \mathcal{C}_{par}} w_c \theta_t^k$.
end for
Return θ_T
Client executes (c, θ_t):
$\theta_t^c \leftarrow \theta_t$ // Initialize local model with global parameters
for $e = 1$ **to** E **do**
  for batch of sample $(X_{1:B}, Y_{1:B} \in \mathcal{D}_c^{\text{ID}})$ **do**
    $\hat{X}_{1:B}$ = Perform Fourier augmentation on $X_{1:B}$.
    $\mathcal{L}_{total} \leftarrow \mathcal{L}_{elbo}^c(\hat{x}, y; \theta_t^c) + \mathcal{L}_{ic}(\hat{x}, y; \theta_t^c)$
    Update θ_t^c using gradient descent on $\mathcal{L}_{total}$
  end for
end for
Return θ_t^c

$k \in \mathcal{C}_{all}$, the following holds:

$$\begin{aligned} &\mathbb{E}_{\tilde{p}(x)} \left| \mathbb{E}_g [y|x] - \tilde{\mathbb{E}} [y|x] \right| \\ &\leq \sigma_\mu^2 \mathbb{E}_{\tilde{p}(x)} \left\| \nabla \sum_{k \in \mathcal{C}_{par}} w_k \log \frac{p_k(f^{-1}(x))}{\tilde{p}(f^{-1}(x))} \right\|_2 \\ &\left\| \mathcal{J}_{f^{-1}(x)} \right\|_2 \|\nabla h\|_2 \big|_{p_k := p_c p_z . (c,z) \sim f^{-1}(x_k)}, \end{aligned} \tag{5}$$

where, $\mathcal{J}_{f^{-1}}s$ represents the Jacobian determinant of f^{-1}. f and h are the mapping functions in the noise addition assumption, where f satisfies bijectivity, h satisfies injectivity, and both are three times continuously differentiable.

Remark 1: This bound indicates that when the global causal mechanism $p(x|z, c)$ is sufficiently strong (i.e., with a smaller σ_μ), it dominates the effect of prior changes, effectively reducing generalization error. The term $\mathbb{E}_{\tilde{p}(x)} \left\| \nabla \sum_{k \in \mathcal{C}_{par}} w_k \log \frac{p_k(f^{-1}(x))}{\tilde{p}(f^{-1}(x))} \right\|_2$ can be interpreted as the Fisher divergence, which measures the difference between each client's prior distribution and the OOD prior. It can also be used to assess the impact of "OOD-ness" on prediction performance. When the prior distribution of some clients results in a smaller Fisher divergence, the generalization error is correspondingly reduced. Furthermore, previous studies have shown that the Fisher divergence is similar in nature to the forward KL divergence and is highly sensitive to regions of the distribution that are not well covered (Durkan and Song 2021). Consequently, when certain clients have uncovered regions, the term $\log(p_k/\tilde{p})$ may approach infinity, leading to an increase in generalization error.

Experiments

Experimental Setups

Datasets. Following SCONE (Bai et al. 2023) and FOOGD (Liao et al. 2024), we select the clear versions of CIFAR-10, CIFAR-100 (Krizhevsky, Hinton et al. 2009), and TinyImageNet (Le and Yang 2015), as ID datasets. To perform OOD generalization, we use corresponding synthetic covariate-shifted datasets as ID-C datasets, which were processed with 15 common image corruption methods. Additionally, we applied 4 extra corruption types to CIFAR-10-C and CIFAR-100-C (Hendrycks and Dietterich 2018). For OOD detection, we chose five external image datasets: LSUN-Crop, LSUN-Resize (Yu et al. 2015) , Textures (Cimpoi et al. 2014), SVHN (Netzer et al. 2011), and iSUN (Xu et al. 2015) to evaluate the model's performance. To evaluate generalization on unseen clients, we use the PACS (Li et al. 2017) dataset, using one domain as an OOD test domain. Dataset simulation details are in Appendix C.

Baseline Methods. We compared FedSDWC with SOTA FL baselines for OOD detection (FedLN (Wei et al. 2022), FOSTER (Yu et al. 2023), FedATOL (Zheng et al. 2023)) and OOD generalization (FedT3A (Iwasawa and Matsuo 2021), FedIIR (Guo et al. 2023), FedTHE (Jiang and Lin 2022), FOOGD (Liao et al. 2024), FedICON (Tan et al. 2023), PerAda (Xie et al. 2024), FedCiR (Li et al. 2024)). The classical FedAvg (McMahan et al. 2017) was also used.

Evaluation Metrics. To evaluate the model's generalization on ID and OOD data, we report the accuracy on ID and OOD data, denoted as **ID-Acc.** and **ID-C-Acc.**, respectively. For OOD detection, we use two metrics: the AUROC (higher is better) and the False Positive Rate at a 95% True Positive Rate (FPR95, lower is better). Details of the experimental setup are available in Appendix C and the open-source code.

Main Results

OOD Generalization. We conducted a comprehensive comparison between the FedSDWC and two types of baselines, performing experiments on CIFAR-10, CIFAR-100, and TinyImageNet datasets. To further verify the generalization ability of the model, we introduced different levels of data contamination under a highly heterogeneous data distribution setting ($\alpha = 0.1$). The results are presented in Table 1, Table D.1, and Table D.2 in Appendix D, respectively. The results show that classic Non-IID algorithms like FedAvg have poor generalization performance, struggling to handle data contamination, especially with significant distribution shifts. While FedIIR attempts to improve this by learning invariant features, its effectiveness remains limited in highly heterogeneous environments. Although FedIIR demonstrates some improvement in stability, its adaptability to OOD samples is still insufficient.

We also found personalized FL models like FedICON and FOSTER to be highly competitive, significantly improving generalization in Non-IID scenarios with data contamination. This suggests that by learning features tailored to each client's specific distribution, personalized approaches can enhance both global model generalization and local performance. In contrast, FedSDWC relaxes the strict assump-

Corruption Type	Method										
	FedAvg	FedLN	FOSTER	FedATOL	FedT3A	FedIIR	FedTHE	FOOGD	PerAda	FedICON	Ours
None	68.03	75.24	90.22	55.93	68.03	68.26	91.05	75.09	82.92	89.06	**94.37**
Brightness	65.44	71.77	88.70	54.44	61.52	66.12	89.71	73.71	79.00	89.18	**93.77**
Spatter	62.18	67.33	85.63	51.54	55.25	60.97	**87.66**	65.31	76.16	85.23	86.79
Gaussian Blur	46.86	55.64	81.88	43.23	48.52	47.51	81.32	53.26	66.95	85.08	**86.26**
Saturate	63.62	71.76	87.06	54.39	58.41	63.32	88.73	71.98	79.73	88.64	**93.29**
Speckle Noise	52.25	54.30	78.63	40.03	48.38	53.20	79.72	57.72	64.09	80.43	**83.55**
Zoom Blur	45.15	54.88	82.41	42.33	49.48	46.57	82.07	52.97	64.80	86.05	**86.87**
Fog	53.89	60.82	83.35	48.17	49.52	54.85	83.35	60.96	66.09	86.35	**90.36**
Shot Noise	52.73	54.55	78.94	39.57	48.82	53.09	80.31	58.31	64.71	80.34	**84.27**
Frosted Glass Blur	43.13	42.33	75.03	28.97	40.41	44.53	**75.42**	45.80	69.26	70.16	73.48
Gaussian Noise	48.66	50.25	76.80	35.20	45.27	49.15	78.37	53.92	61.09	76.82	**82.14**
Motion Blur	41.30	52.65	81.76	41.52	45.51	44.23	80.10	51.05	64.78	81.22	**87.92**
Snow	54.80	60.55	83.05	45.64	51.52	55.52	83.43	61.90	71.56	82.32	**88.59**
Elastic Transform	52.12	61.29	84.82	45.35	52.45	53.21	84.67	59.18	72.58	83.48	**88.24**
Defouce Blur	52.37	61.08	84.44	46.36	52.60	52.72	84.35	58.66	71.58	86.63	**88.08**
Pixelate	56.88	62.00	85.71	46.20	53.34	59.10	84.74	64.37	79.17	85.41	**87.89**
Contrast	41.25	45.02	72.82	38.90	36.93	41.35	71.85	49.14	52.08	**87.09**	81.48
Frost	56.21	58.22	82.76	41.25	52.16	55.91	82.31	63.84	69.26	83.04	**88.49**
Impulse Noise	49.32	50.52	76.52	40.36	42.79	48.45	**78.49**	52.33	59.30	76.43	74.47
Jpeg Compression	61.56	68.61	86.68	47.94	57.64	60.46	87.50	66.55	79.39	86.37	**89.70**
Avg.	53.39	58.94	82.36	44.37	50.92	53.92	82.75	59.80	69.73	83.46	**86.50**

Table 1: The comparison results of federated OOD generalization on Cifar-10 ($\alpha = 0.1$).

tions of invariant learning by incorporating variant features within a novel causal model designed for Non-IID environments. This approach proved superior across all benchmarks (CIFAR-10, CIFAR-100, and TinyImageNet), where it consistently outperformed existing baselines. FedSDWC demonstrated significantly greater robustness and stability, particularly when handling OOD samples, achieving overall SOTA performance in generalization. Fig. 4 shows the average accuracy of various methods on the corrupted CIFAR-10-C and CIFAR-100-C. The results clearly demonstrate that FedSDWC's generalization performance significantly outperforms key baselines such as FedAvg and FOSTER.

OOD Detection Performance. Our method also excels in the OOD detection task under FL. As shown in Table 2, when trained on CIFAR-10 and tested against five external datasets (e.g., LSUN, SVHN), our method consistently outperforms competing approaches like FedAvg and FOSTER, particularly on the key metrics of FPR95 and AUROC. For instance, it achieves the lowest FPR95 values (e.g., 32.61 on LSUN-Crop) and the highest AUROC scores (e.g., 88.82 on LSUN-Crop), demonstrating its superior ability to distinguish between ID and OOD samples.

Ablation Study

Model Stability Analysis. We verify our model's robustness through a statistical analysis of its performance distribution on corrupted data (Fig. 5). Our method excels not only in its central tendency, with a median accuracy of 87.9%, but more critically, in its low dispersion. It achieves an interquartile range (IQR) of 4.8—the smallest among all approaches—indicating that its performance is the most concentrated and consistent across different types of corruption attacks. This demonstrates that our method possesses superior robustness among all evaluated models.

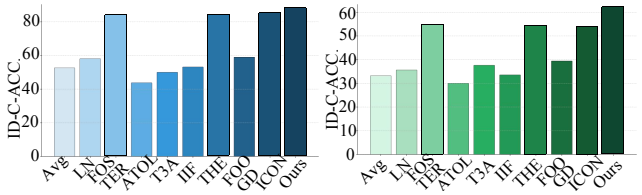


Figure 4: Average Generalization Results of the Model on CIFAR-10-C (left) and CIFAR-100-C (right).

Causal Relationship Analysis Between z and s. We investigate the impact of the causal relationship between z and s on model performance from two perspectives: OOD generalization and detection, as shown in Fig. 7 and Table 3. As shown in Fig. 6, three types of causal relationships exist between z and s: no causal relationship ($z \times s$), causal relationship ($z \rightarrow s$), and weak causal relationship ($z \dashrightarrow s$). The left plot in Fig. 7 shows that regardless of the causal relationship between z and s, the model achieves good performance on the ID data. However, the right plot reveals that the causal relationship between z and s significantly affects the OOD generalization of the model. In particular, the weak causal relationship ($z \dashrightarrow s$) exhibits the best performance, with the accuracy steadily increasing throughout the communication rounds. In contrast, the accuracy of the no causal relationship ($z \times s$) and the causal relationship ($z \rightarrow s$) settings is lower and increases at a slower pace. This phenomenon is further confirmed by Table 3. Weakening the causal rela-

Method	LSUN-Crop		LSUN-Resize		Textures		SVHN		iSUN	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
FedAvg	83.41	58.05	62.01	77.02	80.53	66.23	80.02	62.14	62.10	76.29
FedLN	56.14	84.14	61.31	78.34	93.90	71.99	70.95	76.82	66.41	76.03
FOSTER	47.40	77.43	48.09	76.24	54.23	77.62	39.55	83.07	48.73	76.29
FedATOL	49.50	86.22	64.01	79.89	66.33	78.77	85.39	82.17	61.01	80.05
FedIIR	79.48	63.31	58.44	78.69	91.72	62.32	83.68	64.04	57.86	77.98
FedTHE	58.14	82.04	42.95	83.46	53.58	82.19	39.22	85.95	43.72	83.50
FedICON	48.22	81.28	49.05	83.30	51.57	80.96	**34.94**	85.56	49.98	82.95
PerAda	69.35	71.84	55.77	73.64	76.81	68.25	47.31	78.34	56.97	76.71
Ours	**32.61**	**88.82**	**36.69**	**87.70**	**37.20**	**87.86**	35.46	**88.08**	**36.22**	**87.89**

Table 2: Experimental results of federated OOD detection on Cifar-10 ($\alpha = 0.1$).

Method	LSUN-Crop		LSUN-Resize		Textures		SVHN		iSUN	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
$z \times s$	39.71	86.06	46.35	83.10	44.83	88.15	37.34	86.42	52.92	83.31
$z \rightarrow s$	37.43	86.12	43.14	83.54	34.02	88.66	36.75	87.34	48.42	83.69
$z \dashrightarrow s$	**32.61**	**88.82**	**36.69**	**87.70**	**37.20**	**87.86**	**35.46**	**88.08**	**36.22**	**87.89**

Table 3: Federated OOD detection Under Different Causal Relationships Between z and s (Cifar-10).

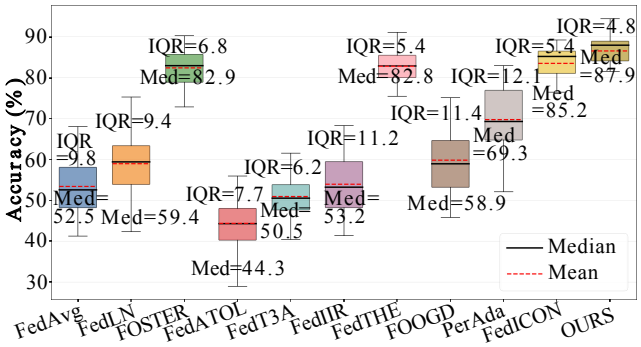


Figure 5: Comparison of anti-corruption performance stability across methods on CIFAR-10-C.

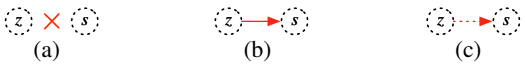


Figure 6: Causal Relationships between z and s.

tionship between z and s yields optimal OOD detection, as measured by FPR95 and AUROC across all datasets.

Visualization. To explore the characteristics of data distribution in FL OOD generalization/detection models, we present the t-SNE visualization of data representations in Fig. 8. The experiments, based on the CIFAR-10 ($\alpha = 5$), compare FedAvg, PerAda, FedIIR, and our FedSDWC. The results show that our method enables a tighter distribution between ID-C data and ID data, while also establishing clearer decision boundaries between ID data and ID-S data.

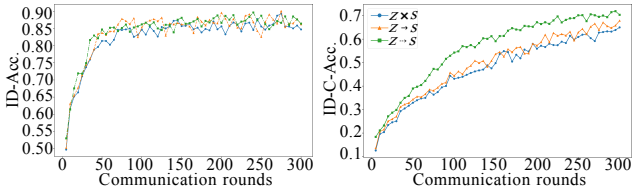
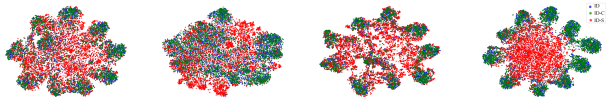


Figure 7: Impact of $z - s$ Causality on Generalization.



(a) FedAvg (b) PerAda (c) FedIIR (d) Ours

Figure 8: T-SNE visualization of different models ($\alpha = 5$).

Conclusion

We proposed FedSDWC, a novel FL model that addresses the dual challenges of OOD generalization and detection. Our model uniquely integrates causal inference with invariant learning by modeling a weak causal relationship between variant and invariant features, operationalized through a novel, intervention-based strategy. This enables the model to leverage useful information from variant features while mitigating their spurious correlations. Theoretically, we provide a generalization error bound and establish the first formal link between FL generalization and client data priors. Empirically, FedSDWC achieves SOTA performance, outperforming existing methods by a significant margin. Future work will focus on enhancing FedSDWC's scalability for massive-scale or extremely heterogeneous data and exploring its application in high-stakes domains such as healthcare.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bai, H.; Canal, G.; Du, X.; Kwon, J.; Nowak, R. D.; and Li, Y. 2023. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *International Conference on Machine Learning*, 1454–1471.

Bai, S.; Zhang, J.; Guo, S.; Li, S.; Guo, J.; Hou, J.; Han, T.; and Lu, X. 2024. DiPrompT: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27284–27293.

Bravo-Hermsdorff, G.; Watson, D.; Yu, J.; Zeitler, J.; and Silva, R. 2024. Intervention generalization: A view from factor graph models. *Advances in Neural Information Processing Systems*, 36.

Chen, H.; Frikha, A.; Krompass, D.; Gu, J.; and Tresp, V. 2023. FRAug: Tackling federated learning with Non-IID features via representation augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4849–4859.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Djurisic, A.; Bozanic, N.; Ashok, A.; and Liu, R. 2023. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *The Eleventh International Conference on Learning Representations*.

Durkan, C.; and Song, Y. 2021. On maximum likelihood training of score-based generative models. *arXiv e-prints*, arXiv–2101.

Galesso, S.; Argus, M.; and Brox, T. 2023. Far Away in the Deep Space: Dense Nearest-Neighbor-Based Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4477–4487.

Gui, S.; Liu, M.; Li, X.; Luo, Y.; and Ji, S. 2024. Joint learning of label and environment causal independence for graph out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36.

Guo, X.; Yu, K.; Cui, L.; Yu, H.; and Li, X. 2025. Federated Causally Invariant Feature Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16978–16986.

Guo, Y.; Guo, K.; Cao, X.; Wu, T.; and Chang, Y. 2023. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, 11905–11933. PMLR.

Hendrycks, D.; and Dietterich, T. G. 2018. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Iwasawa, Y.; and Matsuo, Y. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34: 2427–2440.

Jiang, L.; and Lin, T. 2022. Test-time robust personalization for federated learning. *arXiv preprint arXiv:2205.10920*.

Kong, L.; Xie, S.; Yao, W.; Zheng, Y.; Chen, G.; Stojanov, P.; Akinwande, V.; and Zhang, K. 2022. Partial disentanglement for domain adaptation. In *International conference on machine learning*, 11455–11472. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kumar, S.; Liu, D.; Tian, K.; and Yang, C. 2025. Private Geometric Median in Nearly-Linear Time. *arXiv preprint arXiv:2505.20189*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Lee, R.; Kim, M.; Li, D.; Qiu, X.; Hospedales, T.; Huszár, F.; and Lane, N. 2024. Fedl2p: Federated learning to personalize. *Advances in Neural Information Processing Systems*, 36.

Lee, Y.; Yao, H.; and Finn, C. 2023. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *The Eleventh International Conference on Learning Representations*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Li, H.; Zhang, Z.; Wang, X.; and Zhu, W. 2022. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35: 11828–11841.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, Z.; Lin, Z.; Shao, J.; Mao, Y.; and Zhang, J. 2024. FedCiR: Client-invariant representation learning for federated non-IID features. *IEEE Transactions on Mobile Computing*.

Liao, X.; Liu, W.; Zhou, P.; Yu, F.; Xu, J.; Wang, J.; Wang, W.; Chen, C.; and Zheng, X. 2024. Foogd: Federated collaboration for both out-of-distribution generalization and detection. *Advances in Neural Information Processing Systems*, 37: 132908–132945.

Linderman, R.; Zhang, J.; Inkawhich, N.; Li, H.; and Chen, Y. 2023. Fine-grain inference on out-of-distribution data with hierarchical classification. In *Conference on Lifelong Learning Agents*, 162–183. PMLR.

Liu, C.; Sun, X.; Wang, J.; Tang, H.; Li, T.; Qin, T.; Chen, W.; and Liu, T.-Y. 2021. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34: 6155–6170.

Lv, Z.; Zhang, W.; Zhang, S.; Kuang, K.; Wang, F.; Wang, Y.; Chen, Z.; Shen, T.; Yang, H.; Ooi, B. C.; et al. 2023. Duet: A tuning-free device-cloud collaborative parameters

generation framework for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*, 3077–3085.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Ming, Y.; and Li, Y. 2024. How Does Fine-Tuning Impact Out-of-Distribution Detection for Vision-Language Models? *International Journal of Computer Vision*, 132(2): 596–609.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 5, 7.

Nguyen, T. B.; Nguyen, D. M.; Park, J.; Pham, V. Q.; and Hwang, W.-J. 2025. Federated Domain Generalization with Data-free On-server Matching Gradient. In *The Thirteenth International Conference on Learning Representations*.

Noohdani, F. H.; Hosseini, P.; Parast, A. Y.; Araghi, H. Y.; and Baghshah, M. S. 2024. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27662–27671.

Park, J.; Jung, Y. G.; and Teoh, A. B. J. 2023. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1686–1695.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Qi, Z.; Zhou, S.; Meng, L.; Hu, H.; Yu, H.; and Meng, X. 2025. Federated Deconfounding and Debiasing Learning for Out-of-Distribution Generalization. *arXiv preprint arXiv:2505.04979*.

Sefidgaran, M.; Chor, R.; Zaidi, A.; and Wan, Y. 2024. Lessons from Generalization Error Analysis of Federated Learning: You May Communicate Less Often! In *Forty-first International Conference on Machine Learning*.

Shi, M.; Zhou, Y.; Wang, K.; Zhang, H.; Huang, S.; Ye, Q.; and Lv, J. 2024. PRIOR: Personalized Prior for Reactivating the Information Overlooked in Federated Learning. *Advances in Neural Information Processing Systems*, 36.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840.

Tan, Y.; Chen, C.; Zhuang, W.; Dong, X.; Lyu, L.; and Long, G. 2023. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. *Advances in Neural Information Processing Systems*, 36.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4921–4930.

Wang, K.; Fu, X.; Huang, Y.; Cao, C.; Shi, G.; and Zha, Z.-J. 2023. Generalized uav object detection via frequency domain disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1064–1073.

Wei, H.; Xie, R.; Cheng, H.; Feng, L.; An, B.; and Li, Y. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, 23631–23644. PMLR.

Wu, A.; and Deng, C. 2023. Discriminating known from unknown objects via structure-enhanced recurrent variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23956–23965.

Xie, C.; Huang, D.-A.; Chu, W.; Xu, D.; Xiao, C.; Li, B.; and Anandkumar, A. 2024. PerAda: Parameter-Efficient Federated Learning Personalization with Generalization Guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23838–23848.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14383–14392.

Yang, J.; Zhou, K.; and Liu, Z. 2023. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10): 2607–2622.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, S.; Hong, J.; Wang, H.; Wang, Z.; and Zhou, J. 2023. Turning the curse of heterogeneity in federated learning into a blessing for out-of-distribution detection. In *2023 International Conference on Learning Representations*.

Zhang, J.; Liu, X.; Niu, J.; Tang, S.; Yang, H.; and Wu, X. 2025. Causality Inspired Federated Learning for OOD Generalization. In *Forty-second International Conference on Machine Learning*.

Zhang, R.; Zhou, S.; and Qi, Z. 2025. Federated out-of-distribution generalization: A causal augmentation view. *arXiv preprint arXiv:2504.19882*.

Zheng, H.; Wang, Q.; Fang, Z.; Xia, X.; Liu, F.; Liu, T.; and Han, B. 2023. Out-of-distribution detection learning with unreliable out-of-distribution sources. *Advances in Neural Information Processing Systems*, 36: 72110–72123.

Zheng, Q.; Yang, C.; Yang, H.; and Zhou, J. 2020. A Fast Exact Algorithm for Deployment of Sensor Nodes for Internet of Things. *Information Systems Frontiers*, 22(4): 829–842.

Zhou, P.; Chen, C.; Liu, W.; Liao, X.; Shen, W.; Xu, J.; Fu, Z.; Wang, J.; Wen, W.; and Zheng, X. 2025. FedGOG: Federated Graph Out-of-Distribution Generalization with Diffusion Data Exploration and Latent Embedding Decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22965–22973.