Out-of-Distribution Learning with Human Feedback

Anonymous authors
Paper under double-blind review

Abstract

Out-of-distribution (OOD) learning often relies heavily on statistical approaches or predefined assumptions about OOD data distributions, hindering their efficacy in addressing multifaceted challenges of OOD generalization and OOD detection in real-world deployment environments. This paper presents a novel framework for OOD learning with human feedback, which can provide invaluable insights into the nature of OOD shifts and guide effective model adaptation. Our framework capitalizes on the freely available unlabeled data in the wild that captures the environmental test-time OOD distributions under both covariate and semantic shifts. To harness such data, our key idea is to selectively provide human feedback and label a small number of informative samples from the wild data distribution, which are then used to train a multi-class classifier and an OOD detector. By exploiting human feedback, we enhance the robustness and reliability of machine learning models, equipping them with the capability to handle OOD scenarios with greater precision. We provide theoretical insights on the generalization error bounds to justify our algorithm. Extensive experiments show the superiority of our method, outperforming the current state-of-the-art by a significant margin.

1 Introduction

Modern machine learning models deployed in the wild can inevitably encounter shifts in data distributions. In practice, data shifts can often exhibit heterogeneous forms. For example, out-of-distribution (OOD) data can arise either from semantic shifts where the test data comes from novel categories (Yang et al., 2021), or covariate shifts where the data undergoes domain or environmental changes (Chapaneri & Jayaswal, 2022; Zhou et al., 2022; Koh et al., 2021; Ye et al., 2022). The nature of mixed types of OOD data poses significant challenges in both OOD generalization and OOD detection—necessitating correctly predicting the covariate-shifted OOD samples into one of the known classes, while rejecting the semantic OOD data. For a model to be considered robust and reliable, it must excel in OOD generalization and OOD detection simultaneously, to ensure the continued success and safety of machine learning applications in real-world environments.

Previous works on OOD learning often rely heavily on statistical approaches or predefined assumptions about OOD data distributions, which may not accurately reflect the complexity and diversity of real-world scenarios. Consequently, they struggle to adapt to unforeseen OOD distributions encountered during deployment effectively. Furthermore, without human input to provide contextual information and guide model adaptation, these approaches may face challenges in accurately distinguishing between in-distribution (ID) and OOD data, leading to suboptimal performance in OOD detection tasks. The absence of human feedback thus restricts the adaptability of previous approaches, hindering their efficacy in addressing the multifaceted challenges of OOD generalization and detection in real-world deployment environments.

To tackle the challenge, we introduce a novel framework for OOD learning with human feedback, which can provide invaluable insights into the nature of OOD shifts and guide effective model adaptation. Human feedback offers a unique perspective that complements automated statistical techniques, and remains underexplored in the context of OOD learning. Our framework capitalizes on the abundance of unlabeled data available in the wild, capturing environmental OOD distributions under diverse conditions, which can be characterized as a composite mixture of ID, covariate OOD, and semantic OOD data (Bai et al., 2023). Such unlabeled data is ubiquitous in many real-world applications, arising organically and freely in the model's

operational environment. To harness the unlabeled data for OOD learning, our key idea is to selectively provide human feedback and label a small number of highly informative samples from the wild data distribution, which are then used to train a robust multi-class classifier and a reliable OOD detector. By exploiting human feedback, we can enhance the robustness and reliability of machine learning models, equipping them with the capability to handle OOD scenarios with greater precision.

Our framework employs a gradient-based sample selection mechanism, which prioritizes the most informative samples for human feedback (Section 3.1). The sampling score is calculated based on the projection of its gradient onto the top singular vector of the gradient matrix, defined over all the unlabeled wild data. Specifically, the sampling score measures the norm of the projected vector, which can be used to select informative samples (e.g., ones with relatively large gradient-based scores). The selected samples are then annotated by human, and incorporated into our learning framework. In training, we jointly optimize for both robust classification of samples from the ID and the annotated covariate OOD data, along with a reliable binary OOD detector separating between the ID data and the annotated semantic OOD data (Section 3.2). Additionally, we deliver theoretical insights (Theorem 1) into the learnability of the classifier through the use of a gradient-based sampling score, thus formally justifying the framework of OOD learning with human feedback.

Lastly, we provide extensive experiments showing that this human-centric approach can effectively improve both OOD generalization and detection under a small annotation budget (Section 4). Compared to SCONE (Bai et al., 2023), the current state-of-the-art method, we substantially improve the OOD classification accuracy by 5.82% on covariate-shifted CIFAR-10 data, while reducing the average OOD detection error by 15.16% (FPR95). Moreover, we provide comprehensive ablations on the impacts of labeling budgets, different sampling scores, and sampling strategies, which leads to an improved understanding of OOD learning with human feedback. To summarize our key contributions:

- We propose a new OOD learning with human feedback framework for joint OOD generalization and detection. Our method employs a gradient-based sampling procedure, which can select informative semantic and covariate OOD data from the wild data for OOD learning.
- We present extensive empirical analysis and ablation studies to understand our learning framework. The results provide insights into using human feedback on the unlabeled wild data for both OOD generalization and detection and justify the efficacy of our algorithm.
- We provide a generalization error bound for the model learned under human feedback, theoretically supporting our proposed algorithm.

2 Problem Setup

Labeled in-distribution data. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, ..., C\}$ denote the label space for ID data. We use $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ as the ID joint distribution defined over $\mathcal{X} \times \mathcal{Y}$. Given an ID joint distribution $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$, the labeled ID data $\mathcal{S}^{\text{in}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ is drawn independently and identically (i.i.d.) from $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$.

Unlabeled wild data. Upon deploying a classifier trained on ID, we have access to unlabeled data from the wild, denoted as $\mathcal{S}_{\text{wild}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{m}$, which can be used to assist in OOD learning. Following (Bai et al., 2023), $\mathcal{S}_{\text{wild}}$ is drawn i.i.d. from an unknown wild distribution $\mathbb{P}_{\text{wild}}$ defined below.

Definition 1. *The marginal distribution of the wild data is defined as:*

$$\mathbb{P}_{wild} := (1 - \pi_c - \pi_s)\mathbb{P}_{in} + \pi_c \mathbb{P}_{out}^{covariate} + \pi_s \mathbb{P}_{out}^{semantic},$$

where $\pi_c, \pi_s, \pi_c + \pi_s \in [0, 1]$. $\mathbb{P}_{in}$, $\mathbb{P}_{out}^{covariate}$, *and* $\mathbb{P}_{out}^{semantic}$ *represent the marginal distributions of ID, covariate-shifted OOD, and semantic-shifted OOD data respectively.*

Learning goal. Our learning framework aims to build a robust multi-class predictor $f_{\mathbf{w}}$ and an OOD detector $D_{\boldsymbol{\theta}}$ by leveraging knowledge from labeled ID data $\mathcal{S}^{\text{in}}$ and unlabeled wild data $\mathcal{S}_{\text{wild}}$. Moreover, we allow a maximum number of k human annotations for samples in the unlabeled data. Let $f_{\mathbf{w}} : \mathcal{X} \mapsto \mathbb{R}^C$ be

a multi-class predictor with parameter $\mathbf{w} \in \mathcal{W}$, where $\mathcal{W}$ is the parameter space. The predicted label for an input $\mathbf{x}$ is

$$\widehat{y}(\mathbf{x}; f_{\mathbf{w}}) := \arg\max_{y \in \mathcal{Y}} f_{\mathbf{w}}^{y}(\mathbf{x}),$$

where $f_{\mathbf{w}}^{y}$ is the y-th coordinate of $f_{\mathbf{w}}$, and $\mathbf{x}$ can be either ID or covariate OOD. To detect the semantic OOD data, we need to construct a ranking function $g_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$ with parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter space. With the ranking function $g_{\boldsymbol{\theta}}$, one can define the OOD detector:

$$D_{\boldsymbol{\theta}}(\mathbf{x}; \lambda) := \begin{cases} \text{ID} & \text{if } g_{\boldsymbol{\theta}}(\mathbf{x}) > \lambda, \\ \text{OOD} & \text{if } g_{\boldsymbol{\theta}}(\mathbf{x}) \leq \lambda, \end{cases} \tag{1}$$

where λ is a threshold, typically chosen so that a high fraction of ID data is correctly classified.

3 Proposed Framework

In this section, we introduce a novel framework for OOD learning with human feedback for tackling both OOD generalization and OOD detection problems jointly. Our framework is motivated by the fundamental challenge in harnessing unlabeled wild data for OOD learning—the lack of supervision for samples drawn from the wild data distribution $\mathbb{P}_{\text{wild}}$. To address this challenge, our key idea is to selectively label a small number of samples from the wild data distribution to train a robust multi-class classifier and an OOD detector. Specifically, the design of our framework constitutes two components revolving around the following unexplored questions:

Q1: *How to select informative samples from the unlabeled data for human feedback?* (Section 3.1)

Q2: *How to learn from these newly labeled samples to enhance OOD generalization and OOD detection capabilities?* (Section 3.2)

3.1 Sample Selection for Human Feedback

The key to our OOD learning with human feedback framework lies in a sample selection procedure that identifies the most informative samples while reducing labeling costs. With a limited labeling budget, it is advantageous to select samples from wild data that are either covariate OOD or semantic OOD and will contribute the most to OOD generalization and detection purposes. These samples would be informative for the purpose of OOD generalization and OOD detection. Given a heterogeneous set of wild unlabeled data $\mathcal{S}_{\text{wild}}$, our rationale is to employ a sampling score that can effectively separate ID vs. non-ID part. This way, we can accordingly query samples from the non-ID pool that are most likely covariate or semantic OOD. To achieve this, we proceed to describe the sampling score.

Sampling score. We employ a gradient-based sampling score, where the gradients are estimated from a classification predictor $f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}$ trained on the ID data $\mathcal{S}^{\text{in}}$:

$$\mathbf{w}_{\mathcal{S}^{\text{in}}} \in \arg\min_{\mathbf{w} \in \mathcal{W}} R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}), \tag{2}$$

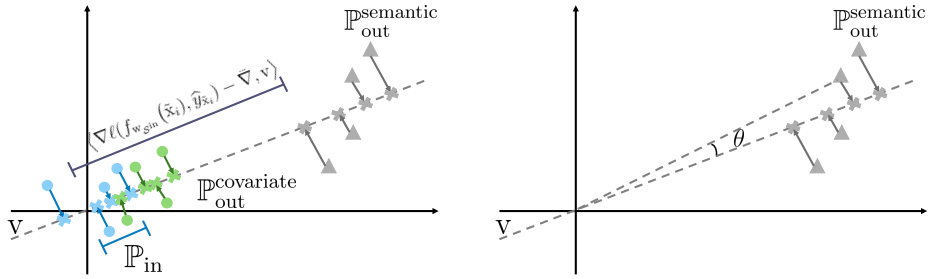
where $R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}) = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{\text{in}}} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$, $\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}^+$ is the loss function, $\mathbf{w}_{\mathcal{S}^{\text{in}}}$ is the learned parameter and n is the size of ID training set. The average gradient $\bar{\nabla}$ is:

$$\bar{\nabla} = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}^{\text{in}}} \nabla \ell(f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\mathbf{x}_i), y_i), \tag{3}$$

where $\bar{\nabla}$ acts as a reference gradient that allows measuring the deviation of any other points from it.

With the reference gradient defined, we can now represent each point in $\mathcal{S}_{\text{wild}}$ as a gradient vector, relative to the reference gradient $\bar{\nabla}$. Specifically, we calculate the gradient matrix (after subtracting the reference gradient $\bar{\nabla}$) for the wild data as follows:

$$\mathbf{G} = \begin{bmatrix} \nabla \ell(f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\tilde{\mathbf{x}}_1), \widehat{y}_{\tilde{\mathbf{x}}_1}) - \bar{\nabla} \\ \cdots \\ \nabla \ell(f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\tilde{\mathbf{x}}_m), \widehat{y}_{\tilde{\mathbf{x}}_m}) - \bar{\nabla} \end{bmatrix}^{\top}, \tag{4}$$



(a) Visualization of the gradient vectors and their projection

(b) Angle between the gradient and the singular vector

Figure 1: Illustration of the gradient vectors and their projections (the blue points denote $\mathbb{P}_{\text{in}}$, the green points represent $\mathbb{P}_{\text{out}}^{\text{covariate}}$, and the gray points indicate $\mathbb{P}_{\text{out}}^{\text{semantic}}$): (a) Visualization of the gradient projected onto the top singular vector of matrix $\mathbf{G}$ for unlabeled data. The gradients of the set $\mathbb{P}_{\text{in}}$ (inliers in the wild) are proximate to the origin (reference gradient $\bar{\nabla}$), in contrast to the gradients of the set $\mathbb{P}_{\text{out}}^{\text{semantic}}$, which are more distant. (b) The angle θ between the gradient of the set $\mathbb{P}_{\text{out}}^{\text{semantic}}$ and the singular vector $\mathbf{v}$. As $\mathbf{v}$ is identified to maximize the distance of the projected points (denoted by ✖) from the origin, considering the sum over all the gradients in $\mathbb{P}_{\text{wild}}$, $\mathbf{v}$ indicates the direction of OOD data in the wild with a small angle θ.

where m denotes the size of the wild dataset, and $\widehat{y}_{\tilde{\mathbf{x}}}$ is the predicted label for a wild sample $\tilde{\mathbf{x}}$.

For each data point $\tilde{\mathbf{x}}_i$ in $\mathcal{S}_{\text{wild}}$, we now define our gradient-based sampling score as follows:

$$\tau_i = \left\langle \nabla \ell(f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\tilde{\mathbf{x}}_i), \widehat{y}_{\tilde{\mathbf{x}}_i}) - \bar{\nabla}, \mathbf{v} \right\rangle^2, \tag{5}$$

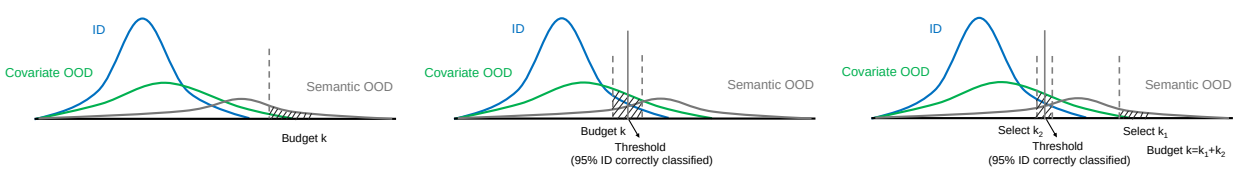
where $\langle \cdot, \cdot \rangle$ is the dot product operator and $\mathbf{v}$ is the top singular vector of $\mathbf{G}$. The top singular vector $\mathbf{v}$ can be regarded as the principal component of the matrix $\mathbf{G}$ in Eq. 4, which maximizes the total distance from the projected gradients (onto the direction of $\mathbf{v}$) to the origin (sum over all points in $\mathcal{S}_{\text{wild}}$) (Hotelling, 1933). Specifically, $\mathbf{v}$ is a unit-norm vector and can be computed as follows:

$$\mathbf{v} \in \arg\max_{\|\mathbf{u}\|_2=1} \sum_{\tilde{\mathbf{x}}_i \in \mathcal{S}_{\text{wild}}} \left\langle \mathbf{u}, \nabla \ell(f_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\tilde{\mathbf{x}}_i), \widehat{y}_{\tilde{\mathbf{x}}_i}) - \bar{\nabla} \right\rangle^2. \tag{6}$$

Essentially, the sampling score τ_i in Eq. 5 measures the ℓ_2 norm of the projected vector. To help readers better understand our design rationale, we provide an illustrative example of the gradient vectors and their projections in Figure 1 (see caption for details).

Sampling strategy. Given the gradient-based scores calculated for each sample $\tilde{\mathbf{x}}_i$ in $\mathcal{S}_{\text{wild}}$, we need to select a subset of $k \ll m$ examples for manual labeling. Here k is the annotation budget. We consider three sampling strategies, as illustrated in Figure 2.

- **Top-k sampling**: select k samples from $\mathcal{S}_{\text{wild}}$ with the largest score τ_i. As shown in Figure 2 (a), these samples deviate mostly from the ID data and are more obviously to be semantic OOD or covariate OOD.



(a) Top-k sampling

(b) Near-boundary sampling

(c) Mixed sampling

Figure 2: Illustration of three selection criteria, (1) top-k sampling, (2) near-boundary sampling, and (3) mixed sampling. The horizontal axis is the sampling score defined in Equation 5, and the vertical axis is the frequency. Note that we color the three different sub-distributions (ID, covariate OOD, semantic OOD) separately for clarity, but in practice, the membership is not revealed due to the unlabeled nature of wild data.

- **Near-boundary sampling**: select k samples that are close to the ID boundary, which may encompass samples with high ambiguity. As shown in Figure 2 (b), we choose the threshold τ_b using labeled ID data $\mathcal{S}^{\text{in}}$ so that it captures a substantial fraction (e.g., 95%) of ID samples. Based on the threshold τ_b, we then select k samples from $\mathcal{S}_{\text{wild}}$ that are closest to this threshold.
- **Mixed sampling**: select samples using both top-k and near-boundary sampling, and combine the two subsets.

Without loss of generality, we denote the selected set of samples as $\mathcal{S}_{\text{selected}}$, with cardinality $|\mathcal{S}_{\text{selected}}| = k$. For each sample in $\mathcal{S}_{\text{selected}}$, we ask the human annotator to choose a label from $\mathcal{Y} \cup \{\perp\}$, where $\{\perp\}$ indicates the semantic OOD. For covariate OOD, the returned label belongs to the existing label space $\mathcal{Y}$ according to the definition. In Section 4.3, we perform comprehensive ablations to understand the efficacy of each sampling strategy.

3.2 Learning Objective Leveraging Human Feedback

We now discuss our learning objective, which incorporates the human annotated samples from wild data. For notation convenience, we use $\mathcal{S}^c_{\text{selected}}$ and $\mathcal{S}^s_{\text{selected}}$ to denote labeled samples corresponding to covariate OOD and semantic OOD respectively, where $\mathcal{S}_{\text{selected}} = \mathcal{S}^c_{\text{selected}} \cup \mathcal{S}^s_{\text{selected}}$. Our learning framework jointly optimizes for both: (1) robust classification of samples from $\mathcal{S}^{\text{in}}$ and covariate OOD $\mathcal{S}^{\text{c}}_{\text{selected}}$, and (2) reliable binary OOD detector separating data between $\mathcal{S}^{\text{in}}$ and semantic OOD $\mathcal{S}^{\text{s}}_{\text{selected}}$. Given a weighting factor α, the risk can be formalized as follows:

$$\mathbf{w}, \boldsymbol{\theta} = \arg\min[\underbrace{R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})}_{\text{Multi-class classifier}} + \alpha \cdot \underbrace{R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{s}}_{\text{selected}}}(g_{\boldsymbol{\theta}})}_{\text{OOD detector}}], \tag{7}$$

where the first term can be empirically optimized using the standard cross-entropy loss. The second term can be viewed as explicitly optimizing the level-set based on the model output (threshold at 0), where the labeled ID data $\mathbf{x}$ from $\mathcal{S}^{\text{in}}$ has positive values and vice versa:

$$\begin{aligned} R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{s}}_{\text{selected}}}(g_{\boldsymbol{\theta}}) &= R^{+}_{\mathcal{S}^{\text{in}}}(g_{\boldsymbol{\theta}}) + R^{-}_{\mathcal{S}^{\text{s}}_{\text{selected}}}(g_{\boldsymbol{\theta}}) \\ &= \mathbb{E}_{\mathbf{x}\in\mathcal{S}^{\text{in}}} \quad \mathbb{1}\{g_{\boldsymbol{\theta}}(\mathbf{x}) \leq 0\} \\ &+ \mathbb{E}_{\tilde{\mathbf{x}}\in\mathcal{S}^{\text{s}}_{\text{selected}}} \quad \mathbb{1}\{g_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) > 0\}. \end{aligned} \tag{8}$$

To make the 0/1 loss tractable, we replace it with the binary sigmoid loss, a smooth approximation of the 0/1 loss. We train $g_{\boldsymbol{\theta}}$ along with the multi-class classifier $f_{\mathbf{w}}$. The training enables generalization to OOD samples drawn from $\mathbb{P}^{\text{covariate}}_{\text{out}}$, and at the same time, teaches the OOD detector to identify data from $\mathbb{P}^{\text{semantic}}_{\text{out}}$. The above process of sample selection, human annotation, and model training can be repeated until the desired performance level is achieved or the entire budget allocated for annotations is exhausted. An end-to-end algorithm is fully specified in Appendix A.

Theoretical insights. We now present theory to support our proposed algorithm. Our main Theorem 1 provides a generalization error bound *w.r.t.* the empirical multi-class classifier $f_{\mathbf{w}}$, learned on ID data and the selected covariate OOD data by the objective $R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})$. We specify several mild assumptions and necessary notations for our theorem in Appendix H. Due to space limitations, we omit unimportant constants and simplify the statements of our theorems. We defer the full formal statements to Appendix I. All proofs can be found in Appendices J and K.

Theorem 1. *(Informal). Let $\mathcal{W}$ be a hypothesis space with a VC-dimension of d. Denote the datasets $\mathcal{S}^{in}$ and $\mathcal{S}^{c}_{selected}$ as the labeled ID and the selected covariate OOD data by active learning, and their sizes are n and m_c, respectively. If $\widehat{\mathbf{w}} \in \mathcal{W}$ minimizes the empirical risk $R_{\mathcal{S}^{in},\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}})$ for classifying the ID and covariate OOD data, and $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}})$, then for any $\delta \in (0,1)$, with probability of at least $1-\delta$, we have*

$$\begin{aligned} R_{\mathbb{P}^{covariate}_{out}}(f_{\widehat{\mathbf{w}}}) \leq R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}^*}) + 2\sup_{\mathbf{w}\in\mathcal{W}} d^{\ell}_{\mathbf{w}}(\mathcal{S}^{in}, \mathcal{S}^{c}_{selected}) \\ + 4\sqrt{\frac{2d\log(2m_c) + \log\frac{2}{\delta}}{m_c}} + 2\gamma + 2\zeta, \end{aligned}$$

where $\zeta = \sqrt{(\frac{1}{n} + \frac{1}{m_c})(\frac{d\log(2n+2m_c)-\log(\delta)}{2})} + M$ *and* $\gamma = \min_{\mathbf{w}\in\mathcal{W}} R_{\mathbb{P}_{in}}(f_{\mathbf{w}})$. *$M$ is the upper bound of the loss function for the multi-class classifier,*

$$d^{\ell}_{\mathbf{w}}(\mathcal{S}^{in}, \mathcal{S}^{c}_{selected}) = \|\nabla R_{\mathcal{S}^{in}}(f_{\mathbf{w}}, \widehat{f}) - \nabla R_{\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}}, \widehat{f})\|_2,$$

where $\widehat{f}$ is a classifier which returns the closest one-hot vector of $f_{\mathbf{w}}$, i.e., $R_{\mathcal{S}^{in}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{in}}\ell(f_{\mathbf{w}}, \widehat{f})$ and $R_{\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{c}_{selected}}\ell(f_{\mathbf{w}}, \widehat{f})$.

Practical implications. Theorem 1 states that the generalization error of the multi-class classifier is upper bounded. If the sizes of the labeled ID n and the selected covariate OOD data m_c are relatively large, the ID loss is small, and the optimal risk on covariate OOD $R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}^*})$ and ID γ is small, then the upper bound will mainly depend on the gradient discrepancy between the ID and covariate OOD data selected by active learning. Notably, the bound synergistically aligns with our gradient-based score (Equation 5). Empirically, we verify these conditions of Theorem 1 and our assumptions in Appendix L, which holds in practice.

4 Experiments

4.1 Settings

Datasets and benchmarks. Following the setup in Bai et al. (2023), we employ CIFAR-10 (Krizhevsky et al., 2009) as $\mathbb{P}_{\text{in}}$ and CIFAR-10-C (Hendrycks & Dietterich, 2018) with Gaussian additive noise as the $\mathbb{P}^{\text{covariate}}_{\text{out}}$. For $\mathbb{P}^{\text{semantic}}_{\text{out}}$, we leverage SVHN (Netzer et al., 2011), Textures (Cimpoi et al., 2014), Places365 (Zhou et al., 2017), and LSUN (Yu et al., 2015). We divide CIFAR-10 training set into 50% labeled as ID and 50% unlabeled. And we mix unlabeled CIFAR-10, CIFAR-10-C, and semantic OOD data to generate the wild dataset. To simulate the wild distribution $\mathbb{P}_{\text{wild}}$, we adopt the same mixture ratio as in SCONE (Bai et al., 2023), where $\pi_c = 0.5$ and $\pi_s = 0.1$. Detailed descriptions of the datasets and data mixture can be found in the Appendix B. To demonstrate the adaptability and robustness of our proposed method, we extend the framework to more diverse settings and datasets. Additional results on other types of covariate shifts can be found in Appendix E.

Experimental details. To ensure a fair comparison with prior works (Bai et al., 2023; Liu et al., 2020; Katz-Samuels et al., 2022), we adopt Wide ResNet with 40 layers and a widen factor of 2 (Zagoruyko & Komodakis, 2016). We use stochastic gradient descent with Nesterov momentum (Duchi et al., 2011), with weight decay 0.0005 and momentum 0.09. The model is initialized with a pre-trained network on CIFAR-10, and then trained for 100 epochs using our objective in Equation 7, with $\alpha = 10$. We use a batch size of 128 and an initial learning rate of 0.1 with cosine learning rate decay. We default k to 1000 and provide analysis of different labeling budgets $k \in \{100, 500, 1000, 2000\}$ in Section 4.3. In our experiment, the output of $g_{\boldsymbol{\theta}}$ is utilized as the score for OOD detection. In practice, we find that using one round of human feedback is sufficient to achieve strong performance. Our implementation is based on PyTorch 1.8.1. All experiments are performed using NVIDIA GeForce RTX 2080 Ti.

Evaluation metrics. We report the accuracy of the ID and covariate OOD data, to measure the classification and OOD generalization performance. In addition, we report false positive rate (FPR) and AUROC for the OOD detection performance. The threshold for the OOD detector is selected based on the ID data when 95% of ID test data points are declared as ID.

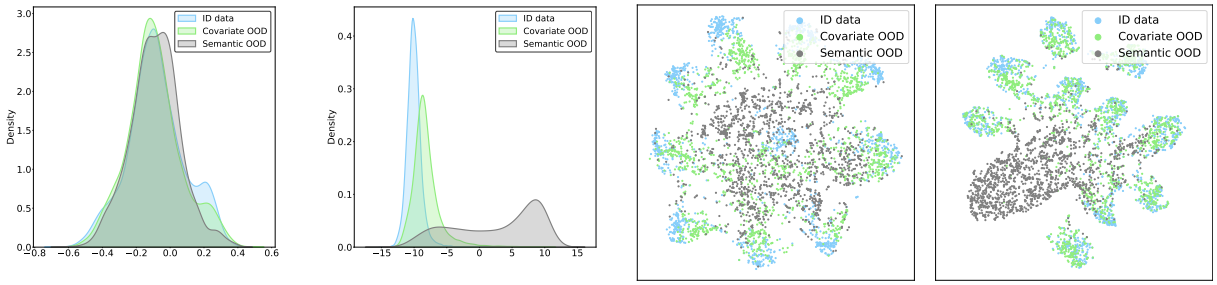
4.2 Main Results

Competitive performance on both OOD detection and generalization tasks. As shown in Table 1, our approach achieves strong performance for both OOD generalization and OOD detection tasks jointly. For a comprehensive evaluation, we compare our method with three categories of methods: **(1)** methods developed specifically for OOD detection, **(2)** methods developed specifically for OOD generalization, and **(3)** methods that are trained with wild data like ours. We discuss them below.

First, we observe that our approach achieves superior performance compared to OOD detection baselines, including MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2018a), Energy (Liu et al., 2020), Maha-

Table 1: **Main results**: comparison with competitive OOD generalization and detection methods on CIFAR-10. Results for LSUN-R and Texture datasets are in Appendix D. *Since all the OOD detection methods use the same model trained with the CE loss on $\mathbb{P}_{\text{in}}$, they display the same ID and OOD accuracy on CIFAR-10-C. We report the average and std of our method based on **3 independent runs**. $\pm x$ denotes the rounded standard error.

Method	SVHN $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				LSUN-C $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				Texture $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$			
	OOD Acc.↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑
OOD detection												
MSP	75.05*	94.84*	48.49	91.89	75.05	94.84	30.80	95.65	75.05	94.84	59.28	88.50
ODIN	75.05	94.84	33.35	91.96	75.05	94.84	15.52	97.04	75.05	94.84	49.12	84.97
Energy	75.05	94.84	35.59	90.96	75.05	94.84	8.26	98.35	75.05	94.84	52.79	85.22
Mahalanobis	75.05	94.84	12.89	97.62	75.05	94.84	39.22	94.15	75.05	94.84	15.00	97.33
ViM	75.05	94.84	21.95	95.48	75.05	94.84	5.90	98.82	75.05	94.84	29.35	93.70
KNN	75.05	94.84	28.92	95.71	75.05	94.84	28.08	95.33	75.05	94.84	39.50	92.73
ASH	75.05	94.84	40.76	90.16	75.05	94.84	2.39	99.35	75.05	94.84	53.37	85.63
OOD generalization												
ERM	75.05	94.84	35.59	90.96	75.05	94.84	8.26	98.35	75.05	94.84	52.79	85.22
Mixup	79.17	93.30	97.33	18.78	79.17	93.30	52.10	76.66	79.17	93.30	58.24	75.70
IRM	77.92	90.85	63.65	90.70	77.92	90.85	36.67	94.22	77.92	90.85	59.42	87.81
VREx	76.90	91.35	55.92	91.22	76.90	91.35	51.50	91.56	76.90	91.35	65.45	85.46
EQRM	75.71	92.93	51.86	90.92	75.71	92.93	21.53	96.49	75.71	92.93	57.18	89.11
SharpDRO	79.03	**94.91**	21.24	96.14	79.03	94.91	5.67	98.71	79.03	**94.91**	42.94	89.99
Learning w. $\mathbb{P}_{wild}$												
OE	37.61	94.68	0.84	99.80	41.37	93.99	3.07	99.26	44.71	92.84	29.36	93.93
Energy (w. outlier)	20.74	90.22	0.86	99.81	32.55	92.97	2.33	99.93	49.34	94.68	16.42	96.46
Woods	52.76	94.86	2.11	99.52	76.90	**95.02**	1.80	99.56	83.14	94.49	39.10	90.45
Scone	84.69	94.65	10.86	97.84	84.58	93.73	10.23	98.02	85.56	93.97	37.15	90.91
Ours	$\mathbf{88.26}_{\pm 0.07}$	$94.68_{\pm 0.07}$	$\mathbf{0.12}_{\pm 0.00}$	$\mathbf{99.98}_{\pm 0.00}$	$\mathbf{90.63}_{\pm 0.02}$	$94.33_{\pm 0.01}$	$\mathbf{0.07}_{\pm 0.00}$	$\mathbf{99.97}_{\pm 0.00}$	$\mathbf{90.31}_{\pm 0.02}$	$94.33_{\pm 0.03}$	$\mathbf{4.91}_{\pm 0.03}$	$\mathbf{98.28}_{\pm 0.01}$



(a) Scores of ERM (b) Scores of Ours (c) T-SNE of ERM (d) T-SNE of Ours

Figure 3: (a)-(b) Score distributions for ERM vs. our method. Different colors represent the different types of test data: CIFAR-10 as $\mathbb{P}_{\text{in}}$ (blue), CIFAR-10-C as $\mathbb{P}_{\text{out}}^{\text{covariate}}$ (green), and Textures as $\mathbb{P}_{\text{out}}^{\text{semantic}}$ (gray). (c)-(d): T-SNE visualization of the image embeddings using ERM vs. our method.

lanobis (Lee et al., 2018), ViM (Wang et al., 2022b), KNN (Sun et al., 2022), and latest baseline ASH (Djurisic et al., 2023a). Methods tailored for OOD detection tend to capture the domain-variant information and struggle with the covariate distribution shift, resulting in suboptimal OOD accuracy. For example, our method achieves near-perfect FPR95 (0.12%), when evaluating against SVHN as semantic OOD. Secondly, while approaches for OOD generalization, containing IRM (Arjovsky et al., 2019), ERM, Mixup (Zhang et al., 2018), VREx (Krueger et al., 2021), EQRM (Eastwood et al., 2022a), and latest baseline SharpDRO (Huang et al., 2023b), demonstrate improved OOD accuracy, they cannot effectively distinguish between ID data and semantic OOD data, leading to poor OOD detection performance. Lastly, closest to our setting, we compare with strong baselines trained with wild data, namely Outlier Exposure (Hendrycks et al., 2018), Energy-regularized learning (Liu et al., 2020), WOODS (Katz-Samuels et al., 2022), and SCONE (Bai et al., 2023). These methods emerge as robust OOD detectors, yet display a notable decline in OOD generalization (except for SCONE). In contrast, our method demonstrates consistently better results in terms of both OOD generalization and detection performance. Notably, our method even surpasses the current state-of-the-art method SCONE by 32.24% in terms of FPR95 on the Texture OOD dataset, and simultaneously improves the OOD accuracy by 4.75% on CIFAR-10-C.

Additional results on PACS benchmark. In Table 2, we report results on the PACS dataset (Li et al., 2017) from DomainBed. We compare our approach with various common OOD generalization baselines, including IRM (Arjovsky et al., 2019), DANN (Ganin et al., 2016), CDANN (Li et al., 2018c), GroupDRO (Sagawa et al., 2020), MTL (Blanchard et al., 2021), I-Mixup (Wang et al., 2020), MMD (Li et al., 2018b), VREx (Krueger et al., 2021), MLDG (Li et al., 2018a), ARM (Zhang et al., 2021b), RSC (Huang et al., 2020), Mixstyle (Zhou et al., 2021), ERM (Vapnik, 1999), CORAL (Sun & Saenko, 2016), SagNet (Nam

Algorithm	Art painting	Cartoon	Photo	Sketch	Avg (%)
IRM (Arjovsky et al., 2019)	84.8	76.4	96.7	76.1	83.5
DANN (Ganin et al., 2016)	86.4	77.4	97.3	73.5	83.7
CDANN (Li et al., 2018c)	84.6	75.5	96.8	73.5	82.6
GroupDRO (Sagawa et al., 2020)	83.5	79.1	96.7	78.3	84.4
MTL (Blanchard et al., 2021)	87.5	77.1	96.4	77.3	84.6
I-Mixup (Wang et al., 2020)	86.1	78.9	97.6	75.8	84.6
MMD (Li et al., 2018b)	86.1	79.4	96.6	76.5	84.7
VREx (Krueger et al., 2021)	86.0	79.1	96.9	77.7	84.9
MLDG (Li et al., 2018a)	85.5	80.1	97.4	76.6	84.9
ARM (Zhang et al., 2021b)	86.8	76.8	97.4	79.3	85.1
RSC (Huang et al., 2020)	85.4	79.7	97.6	78.2	85.2
Mixstyle (Zhou et al., 2021)	86.8	79.0	96.6	78.5	85.2
ERM (Vapnik, 1999)	84.7	80.8	97.2	79.3	85.5
CORAL (Sun & Saenko, 2016)	88.3	80.0	97.5	78.8	86.2
SagNet (Nam et al., 2021)	87.4	80.7	97.1	80.0	86.3
SelfReg (Kim et al., 2021)	87.9	79.4	96.8	78.3	85.6
GVRT Min et al. (2022)	87.9	78.4	98.2	75.7	85.1
VNE (Kim et al., 2023)	**88.6**	79.9	96.7	82.3	86.9
Ours	88.1	**87.4**	**98.5**	**91.3**	**91.3**

Table 2: Comparison with domain generalization methods on the PACS benchmark. All methods are trained on ResNet-50. The model selection is based on a training domain validation set.

et al., 2021), SelfReg (Kim et al., 2021), GVRT (Min et al., 2022), and the latest baseline VNE (Kim et al., 2023). Our method achieves an average accuracy of 91.3%, which outperforms these OOD generalization baselines.

Visualization of OOD score distributions. Figure 3 (a) and (b) visualize the score distributions for ERM (without human feedback) vs. our method. The OOD score distributions between $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}^{\text{semantic}}$ are more clearly differentiated using our method. This separation leads to an improvement in OOD detection performance. The enhanced separation can be attributed to the effectiveness of the OOD learning with human feedback framework in recognizing semantic-shifted OOD data.

Visualization of feature embeddings. Figure 3 (c) and (d) present t-SNE visualizations (Van der Maaten & Hinton, 2008) of feature embeddings on the test data. The blue points represent the test ID data (CIFAR-10), green points denote test samples from CIFAR-10-C, and gray points are from the Texture dataset. This visualization indicates that embeddings of $\mathbb{P}_{\text{in}}$ (CIFAR) and $\mathbb{P}_{\text{out}}^{\text{covariate}}$ (CIFAR-C) are more closely aligned using our method, which contributes to enhanced OOD generalization performance.

4.3 In-Depth Analysis

Impact of labeling budget k. The budget k is central to our OOD learning with human feedback framework. In table 3, we conduct an ablation by varying $k \in \{100, 500, 1000, 2000\}$. We observe that the performance of OOD generalization and OOD detection both increase with a larger number of annotation budgets. For example, The OOD accuracy improves from 86.62% to 91.09% when the budget changes from $k = 100$ to $k = 2000$. At the same time, the FPR95 reduces from 22.16% to 3.60%. Interestingly, we do notice a marginal difference between $k = 1000$ and $k = 2000$, which suggests that our method suffices to achieve strong performance without excessive labeling budget.

Table 3: Ablation on labeling budget k. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Budget k	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑
100	86.62	95.03	22.16	91.34
500	89.96	94.70	7.50	97.45
1000	90.31	94.33	4.91	98.28
2000	91.09	94.42	3.60	98.90

Impact of different sampling strategies. In Table 4, we compare the performance of using three different sampling strategies (1) top-k sampling, (2) near-boundary sampling, and (3) mixed sampling. For all three strategies, we employ the same labeling budget $k = 100$ and the same gradient-based scoring function (*c.f.* Section 3.1). We observe that the top-k sampling achieves the best OOD generalization

performance. This is because the selected samples are furthest away from the ID data, presenting challenging cases of covariate-shifted OOD data. By labeling and learning from these hard cases, the multi-class classifier acquires a stronger generalization to OOD data. Moreover, near-boundary sampling displays the lowest performance in both OOD generalization and OOD detection. To reason for this, we provide the number of samples (among 100) belonging to ID, covariate OOD, and semantic OOD respectively. As seen in Table 4, the majority of the samples appear to be either ID or covariate OOD (easy cases), whereas only 6 out of 100 samples are semantic OOD. As a result, this sampling strategy does not provide sufficient informative samples needed for the OOD detector. Lastly, the mixed sampling strategy achieves performance somewhere in between top-k sampling and near-boundary sampling, which aligns with our expectations.

Table 4: Impact of sampling strategy ($k = 100$). We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Sampling Strategies	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑	#ID	#Covariate OOD	#Semantic OOD
Top-k sampling	86.62	95.03	22.16	91.34	0	57	43
Near-boundary sampling	85.12	95.18	41.72	76.56	44	50	6
Mixed sampling	85.36	95.10	36.24	85.37	25	51	24

Impact of different sampling scores. Our main results are based on the gradient-based sampling score (*c.f.* Section 3.1). Here we provide additional comparison using different sampling scores, including least-confidence sampling Wang & Shang (2014), entropy-based sampling Wang & Shang (2014), margin-based sampling Roth & Small (2006), energy score Liu et al. (2020), BADGE Ash et al. (2019), and random sampling. Detailed description for each method is provided in Appendix C. We observe that the gradient-based score demonstrates overall strong performance in terms of OOD generalization and detection.

Table 5: Impact of sampling scores. We use budget $k = 1000$ for all methods. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Sampling score	OOD Acc.↑	ID Acc.↑	FPR↓	AUROC↑
Least confidence	90.22	94.73	8.44	96.45
Entropy	90.02	94.80	8.19	96.66
Margin	90.31	94.56	6.37	97.68
BADGE	88.68	94.56	8.77	96.39
Random	89.22	**94.84**	9.45	95.41
Energy score	88.75	94.78	10.89	95.19
Gradient-based	**90.31**	94.33	**4.91**	**98.28**

5 Related Works

Out-of-distribution generalization is a crucial problem in machine learning when training and test data are sampled from different distributions. Compared to domain adaptation task (You et al., 2019; Kumar et al., 2020; Wang et al., 2022c; Prabhu et al., 2021; Su et al., 2020; Kothandaraman et al., 2023), OOD generalization is more challenging, as it focuses on adapting to *unseen* covariate-shifted data without access to any sample from the target domain (Gulrajani & Lopez-Paz, 2020; Bai et al., 2021; Koh et al., 2021; Ye et al., 2022; Cho et al., 2023). Existing theoretical work on domain adaptation (Mansour et al., 2009; Blitzer et al., 2007; Ben-David et al., 2006; 2010; Gui et al., 2024) discusses the generalization error bound with access to target domain data. Our analysis presents several key distinctions: (1) Our focus is on the wild setting, necessitating an additional step of selection and human annotation to acquire selected covariate OOD data for retraining; (2) Our OOD generalization error bound differs in that it is based on a gradient-based discrepancy between ID and OOD data. This diverges from classical domain adaptation literature and synergistically aligns with our gradient-based sampling score.

A prevalent approach in the OOD generalization area is to learn a domain-invariant data representation across training domains. This involves various strategies like invariant risk minimization (Arjovsky et al., 2019; Ahuja et al., 2020; Krueger et al., 2021; Eastwood et al., 2022b), robust optimization (Sagawa et al., 2020; Dai et al., 2023; Huang et al., 2023a), domain adversarial learning (Li et al., 2018b; Wang et al., 2022d; Dayal et al., 2023), meta-learning (Li et al., 2018a; Zhang et al., 2021a), and gradient alignment (Shi et al., 2021; Rame et al., 2022; Guo et al., 2023). Some OOD algorithms do not require multiple training domains (Tong et al., 2023). Other approaches include model ensembles (Chen et al., 2023c; Ramé et al., 2023), graph learning (Gui et al., 2023; Yuan et al., 2023), and test-time adaptation (Chen et al., 2023a; Park et al., 2023; Samadh et al., 2023; Chen et al., 2023b). SCONE (Bai et al., 2023) aims to enhance OOD

robustness and detection by utilizing wild data from the open world. Building on SCONE's problem setting, we leverage human feedback to train a robust classifier and OOD detector, supported by theoretical analysis.

Out-of-distribution detection has garnered increasing interest in recent years (Yang et al., 2021). Recent methods can be broadly classified into post hoc and regularized-based algorithms. Post hoc methods, which include confidence-based methods (Hendrycks & Gimpel, 2016; Liang et al., 2018b), energy-based scores (Liu et al., 2020; Wang et al., 2021; Sun et al., 2021; Djurisic et al., 2023b; Zhang et al., 2023b; Lafon et al., 2023), gradient-based score (Huang et al., 2021; Behpour et al., 2023), Bayesian approaches (Gal & Ghahramani, 2016; Maddox et al., 2019; Kristiadi et al., 2020), and distance-based methods (Lee et al., 2018; Sun et al., 2022; Du et al., 2022; Ming et al., 2022b), perform OOD detection by devising OOD scores at test time. On the other hand, another line of work addresses OOD detection through training-time regularization (Hendrycks et al., 2018; Hein et al., 2019; Ming et al., 2022a; Wang et al., 2023), which typically relies on a clean set of auxiliary semantic OOD data. WOODS (Katz-Samuels et al., 2022) and SAL (Du et al., 2024) address this issue by utilizing wild mixture data, comprising both unlabeled ID and semantic OOD data. While SAL also employs a gradient-based score, it does not consider OOD generalization or leveraging human feedback, which is our main focus. Our work builds upon the setting in SCONE (Bai et al., 2023) and introduces a novel OOD learning with human feedback framework aimed at enhancing both OOD generalization and detection jointly.

Learning with human feedback emphasizes the selection of the most informative data points for labeling (Cohn et al., 1994; Balcan et al., 2006; Settles, 2009; Wang & Shang, 2014; Ren et al., 2021; Karzand & Nowak, 2020). Well-known sampling strategies include disagreement-based sampling, diversity sampling, and uncertainty sampling. Disagreement-based sampling (Seung et al., 1992; Hanneke et al., 2014; Zhu & Nowak, 2022) focuses on selecting data points that elicit disagreement among multiple models. Diversity sampling, as explored in (Du et al., 2015; Zhdanov, 2019; Citovsky et al., 2021), aims to select data points that are both diverse and representative of the data's overall distribution. Uncertainty sampling (Lewis, 1995; Scheffer et al., 2001; Shannon, 2001; Lu et al., 2016; Wang & Shang, 2014; Ducoffe & Precioso, 2018; Beluch et al., 2018) seeks to identify data points where model confidence is lowest, thus reducing uncertainty. More advanced methods (Ash et al., 2019; Wang et al., 2022a; Elenter et al., 2022; Mohamadi et al., 2022) incorporate a mix of uncertainty and diversity sampling techniques. Another research direction involves deep active learning under data imbalance (Kothawade et al., 2021; Emam et al., 2021; Zhang et al., 2022; Coleman et al., 2022; Zhang et al., 2023a; Kothawade et al., 2022; Aggarwal et al., 2020). The work of (Das et al., 2023; Deng et al., 2023; Zhan et al., 2023; Shayovitz et al., 2024; Benkert et al., 2022) considers distribution shifts in the context of active learning. However, previous approaches do not consider OOD robustness and the challenges posed by realistic scenarios involving wild data. In our work, we introduce a novel framework tailored for both OOD generalization and detection challenges. This framework is further supported by a theoretical justification of our learning approach.

6 Conclusion

This paper introduces a new framework leveraging human feedback for both OOD generalization and OOD detection. Our framework tackles the fundamental challenge of leveraging the wild data—the lack of supervision for samples from the wild distribution. Specifically, we employ a gradient-based sampling score to selectively label informative OOD samples from the wild data distribution and then train a robust multi-class classifier and an OOD detector. Importantly, our algorithm only requires a small annotation budget and performs competitively compared to various baselines, which offers practical advantages. We further provide theoretical analysis of the learnability of the classifier. We hope our work will inspire future research on both empirical and theoretical understanding of OOD generalization and detection in a synergistic way.

References

Umang Aggarwal, Adrian Popescu, and Céline Hudelot. Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1428–1437, 2020.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pp. 145–155. PMLR, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2019.

Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6705–6713, 2021.

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

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 65–72, 2006.

Sima Behpour, Thang Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Advances in Neural Information Processing Systems*, 2023.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Ryan Benkert, Mohit Prabhushankar, and Ghassan AlRegib. Forgetful active learning with switch events: Efficient sampling for out-of-distribution data. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2196–2200. IEEE, 2022.

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007.

Santosh Chapaneri and Deepak Jayaswal. Covariate shift in machine learning. In *Handbook of Research on Machine Learning*, pp. 87–119. Apple Academic Press, 2022.

Chaoqi Chen, Luyao Tang, Yue Huang, Xiaoguang Han, and Yizhou Yu. Coda: Generalizing to open and unseen domains with compaction and disambiguation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24172–24182, 2023b.

Yimeng Chen, Tianyang Hu, Fengwei Zhou, Zhenguo Li, and Zhi-Ming Ma. Explore and exploit the diverse knowledge in model zoo for domain generalization. In *International Conference on Machine Learning*, pp. 4623–4640. PMLR, 2023c.

Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15702–15712, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

Cody Coleman, Edward Chou, Julian Katz-Samuels, Sean Culatana, Peter Bailis, Alexander C Berg, Robert Nowak, Roshan Sumbaly, Matei Zaharia, and I Zeki Yalniz. Similarity search for efficient active learning and search of rare concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6402–6410, 2022.

Rui Dai, Yonggang Zhang, Zhen Fang, Bo Han, and Xinmei Tian. Moderately distributional exploration for domain generalization. 2023.

Arnav Mohanty Das, Gantavya Bhatt, Megh Manoj Bhalerao, Vianne R Gao, Rui Yang, and Jeff Bilmes. Accelerating batch active learning using continual learning techniques. *Transactions on Machine Learning Research*, 2023.

Aveen Dayal, KB Vimal, Linga Reddy Cenkeramaddi, C Krishna Mohan, Abhinav Kumar, and Vineeth N Balasubramanian. Madg: Margin-based adversarial learning for domain generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. Counterfactual active learning for out-of-distribution generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11362–11377, 2023.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*. OpenReview.net, 2023a.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023b.

Bo Du, Zengmao Wang, Lefei Zhang, Liangpei Zhang, Wei Liu, Jialie Shen, and Dacheng Tao. Exploring representativeness and informativeness for active learning. *IEEE transactions on cybernetics*, 47(1):14–26, 2015.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. *Advances in Neural Information Processing Systems*, 35:20434–20449, 2022.

Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations*, 2024.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. In *NeurIPS*, 2022a.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022b.

Juan Elenter, Navid NaderiAlizadeh, and Alejandro Ribeiro. A lagrangian duality approach to active learning. *Advances in Neural Information Processing Systems*, 35:37575–37589, 2022.

Zeyad Ali Sami Emam, Hong-Min Chu, Ping-Yeh Chiang, Wojciech Czaja, Richard Leapman, Micah Goldblum, and Tom Goldstein. Active learning at the imagenet scale. *arXiv preprint arXiv:2111.12880*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. 2023.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. *arXiv preprint arXiv:2404.05094*, 2024.

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

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, pp. 11905–11933. PMLR, 2023.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pp. 124–140, 2020.

Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, and Tongliang Liu. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16175–16185, 2023a.

Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, and Tongliang Liu. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *CVPR*, pp. 16175–16185. IEEE, 2023b.

Mina Karzand and Robert D Nowak. Maximin active learning in overparameterized model classes. *IEEE Journal on Selected Areas in Information Theory*, 1(1):167–177, 2020.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pp. 180–191. Toronto, Canada, 2004.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *IEEE International Conference on Computer Vision*, pp. 9619–9628, 2021.

Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3799–3810, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Divya Kothandaraman, Sumit Shekhar, Abhilasha Sancheti, Manoj Ghuhan, Tripti Shukla, and Dinesh Manocha. Salad: Source-free active label-agnostic domain adaptation for classification, segmentation and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 382–391, 2023.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

Suraj Kothawade, Shivang Chopra, Saikat Ghosh, and Rishabh Iyer. Active data discovery: Mining unknown data using submodular information measures. *arXiv preprint arXiv:2206.08566*, 2022.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *International conference on machine learning*, pp. 5436–5446. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.

Marc Lafon, Elias Ramzi, Clément Rambour, and Nicolas Thome. Hybrid energy based model in the feature space for out-of-distribution detection. In *International Conference on Machine Learning*, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Yunwen Lei and Yiming Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 2021.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*, pp. 624–639, 2018c.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018a.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018b.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Jing Lu, Peilin Zhao, and Steven CH Hoi. Online passive-aggressive active learning. *Machine Learning*, 103: 141–183, 2016.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim. Grounding visual representations with texts for domain generalization. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2022.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665. PMLR, 2022a.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2022b.

Mohamad Amin Mohamadi, Wonho Bae, and Danica J Sutherland. Making look-ahead active learning strategies feasible with neural tangent kernels. *Advances in Neural Information Processing Systems*, 35: 12542–12553, 2022.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Jungwuk Park, Dong-Jun Han, Soyeong Kim, and Jaekyun Moon. Test-time style shifting: Handling arbitrary styles in domain generalization. 2023.

Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8505–8514, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 18347–18377. PMLR, 2022.

Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pp. 28656–28679. PMLR, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *Machine Learning: ECML 2006: 17th European Conference on Machine Learning Berlin, Germany, September 18-22, 2006 Proceedings 17*, pp. 413–424. Springer, 2006.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International symposium on intelligent data analysis*, pp. 309–318. Springer, 2001.

Burr Settles. Active learning literature survey. 2009.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

Shachar Shayovitz, Koby Bibas, and Meir Feder. Deep individual active learning: Safeguarding against out-of-distribution challenges in neural networks. *Entropy*, 26(2):129, 2024.

Yuge Shi, Jeffrey Seely, Philip Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2021.

Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 739–748, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450, 2016.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Peifeng Tong, Wu Su, He Li, Jialin Ding, Zhan Haoxiang, and Song Xi Chen. Distribution free domain generalization. In *International Conference on Machine Learning*, pp. 34369–34378. PMLR, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10 (5):988–999, 1999.

Dan Wang and Yi Shang. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, pp. 112–119. IEEE, 2014.

Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, 35:25171–25184, 2022a.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022b.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022c.

Qixun Wang, Yifei Wang, Hong Zhu, and Yisen Wang. Improving out-of-distribution generalization by adversarial training with structured priors. *Advances in Neural Information Processing Systems*, 35: 27140–27152, 2022d.

Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3622–3626, 2020.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7947–7958, 2022.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2720–2729, 2019.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. Environment-aware dynamic graph learning for out-of-distribution generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Xueying Zhan, Zeyu Dai, Qingzhong Wang, Qing Li, Haoyi Xiong, Dejing Dou, and Antoni B Chan. Pareto optimization for active learning under out-of-distribution data scenarios. *Transactions on Machine Learning Research*, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Jifan Zhang, Julian Katz-Samuels, and Robert Nowak. Galaxy: Graph-based active learning at the extreme. In *International Conference on Machine Learning*, pp. 26223–26238. PMLR, 2022.

Jifan Zhang, Shuai Shao, Saurabh Verma, and Robert Nowak. Algorithm selection for deep active learning with imbalanced datasets. *arXiv preprint arXiv:2302.07317*, 2023a.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2023b.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021a.

Marvin Mengxin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, 2021b.

Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464, 2017.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yinglun Zhu and Robert Nowak. Active learning with neural networks: Insights from nonparametric statistics. *Advances in Neural Information Processing Systems*, 35:142–155, 2022.

Out-of-Distribution Learning with Human Feedback (Appendix)

A Algorithm

Algorithm 1 Out-of-Distribution Learning with Human Feedback

Input: In-distribution labeled data $\mathcal{S}^{\text{in}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Unlabeled wild data $\mathcal{S}_{\text{wild}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^m$.
Output: Learned classifier $f_{\widehat{\mathbf{w}}}$ and OOD detector $g_{\widehat{\boldsymbol{\theta}}}$.
Sample section
1. Perform ERM on labeled ID data $\mathcal{S}^{\text{in}}$ and obtain learned weight parameter $\mathbf{w}_{\mathcal{S}^{\text{in}}}$ according to Eq. equation 2.
2. Calculate the reference gradient $\bar{\nabla}$ on $\mathcal{S}^{\text{in}}$ according to Eq. equation 3.
3. Generate predicted labels $\widehat{y}_{\tilde{\mathbf{x}}_i}$ for $\tilde{\mathbf{x}}_i \in \mathcal{S}_{\text{wild}}$.
4. Calculate gradient $\nabla \ell(h_{\mathbf{w}_{\mathcal{S}^{\text{in}}}}(\tilde{\mathbf{x}}_i), \widehat{y}_{\tilde{\mathbf{x}}_i})$.
5. Calculate the gradient matrix $\boldsymbol{G}$ by Eq. equation 4 and Compute the gradient-based score τ_i by Eq. equation 5.
6. Given the gradient-based scores τ_i for each sample $\tilde{\mathbf{x}}_i$, select a subset k according to the three sampling strategies described in Sec. 3.1.
7.Annotate the selected k samples with labels from: $\mathcal{Y} \cup \{\perp\}$ to obtain $\mathcal{S}^{\text{c}}_{\text{selected}}$ and $\mathcal{S}^{\text{s}}_{\text{selected}}$, where $\{\perp\}$ indicates the semantic OOD.
OOD learning with annotated samples
8. Train the robust classifier using samples from $\mathcal{S}^{\text{in}}$ and covariate OOD $\mathcal{S}^{\text{c}}_{\text{selected}}$. Concurrently, train a binary OOD detector using semantic OOD $\mathcal{S}^{\text{s}}_{\text{selected}}$ and $\mathcal{S}^{\text{in}}$ by Eq. equation 7.

B Detailed Description of Datasets

CIFAR-10 (Krizhevsky et al., 2009) contains $60,000$ color images with 10 classes. The training set has $50,000$ images and the test set has $10,000$ images.

CIFAR-10-C is algorithmically generated, following the previous leterature (Hendrycks & Dietterich, 2018). The corruptions include Gaussian noise, defocus blur, glass blur, impulse noise, shot noise, snow, and zoom blur.

SVHN (Netzer et al., 2011) is a real-world image dataset obtained from house numbers in Google Street View images, with 10 classes. This dataset contains $73,257$ samples for training and $26,032$ samples for testing.

Places365 (Zhou et al., 2017) contains scene photographs and diverse types of environments encountered in the world. The scene semantic categories consist of three macro-classes: Indoor, Nature, and Urban.

LSUN-C (Yu et al., 2015) and **LSUN-R** (Yu et al., 2015) are large-scale image datasets that are annotated using deep learning with humans in the loop. LSUN-C is a cropped version of LSUN and LSUN-R is a resized version of the LSUN dataset.

Textures (Cimpoi et al., 2014) refers to the Describable Textures Dataset, which contains images of patterns and textures. The subset we used has no overlap categories with the CIFAR dataset (Krizhevsky et al., 2009).

PACS (Li et al., 2017) is commonly used in evaluating OOD generalization approaches. This dataset consists of $9,991$ examples of resolution 224×224 and four domains with different image styles including photo, art painting, cartoon, and sketch with seven categories.

Details of data split for OOD datasets. For datasets with standard train-test split (e.g., SVHN), we use the original test split for evaluation. For other OOD datasets (e.g., LSUN-C), we use 70% of the data for creating the wild mixture training data as well as the mixture validation dataset. We use the

remaining examples for test-time evaluation. For splitting training/validation, we use 30% for validation and the remaining for training.

C Description of Sampling Methods

Least confidence (Wang & Shang, 2014) is an uncertainty-based algorithm that selects data for which the most probable label possesses the lowest posterior probability. This method focuses on instances where the model's predictions are least certain.

Margin sampling (Roth & Small, 2006) elects data points based on the multiclass margin, specifically targeting examples where the posterior probabilities of the two most likely labels are closely matched, indicating the minimal difference between them.

Entropy (Wang & Shang, 2014) is an uncertainty-based algorithm that chooses data points by evaluating the entropy within the predictive class probability distribution of each example, aiming to maximize the overall predictive entropy.

Energy sampling (Liu et al., 2020) identifies data points using an energy score, a measure theoretically aligned with the probability density of the inputs.

BADGE (Ash et al., 2019) samples groups of points that are both diverse and exhibit high magnitude in a hallucinated gradient space. This technique combines predictive uncertainty and sample diversity, enhancing the effectiveness of data selection in active learning.

Random sampling serves as a straightforward baseline method, involving the random selection of k examples to query.

D Results on Additional OOD Datasets

In this section, we provide the main results on more OOD datasets including Places365 (Zhou et al., 2017) and LSUN-R (Yu et al., 2015). As shown in Table 6. our proposed approach achieves overall strong performance in OOD generalization and OOD detection on these additional OOD datasets. Firstly, we compare our method with post hoc OOD detection methods such as MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2018a), Energy (Liu et al., 2020), Mahalanobis (Lee et al., 2018), ViM (Wang et al., 2022b), KNN (Sun et al., 2022), and latest baseline ASH (Djurisic et al., 2023a). These methods are all based on a model trained with cross-entropy loss, which suffers from limiting OOD generalization performance. However, our method achieves an improved OOD generalization performance (e.g., 91.08% when the wild data is a mixture of CIFAR-10, CIFAR-10-C, and LSUN-R).

Secondly, we also compare our method with common OOD generalization baseline methods including IRM (Arjovsky et al., 2019), ERM, Mixup (Zhang et al., 2018), VREx (Krueger et al., 2021), EQRM (Eastwood et al., 2022a), and latest baseline SharpDRO (Huang et al., 2023b). Our approach consistently achieves better results compared to these OOD generalization baselines. Lastly, we compare our method with strong OOD baselines using $\mathbb{P}_{\text{wild}}$ such as Outlier Exposure (Hendrycks et al., 2018), Energy-regularized learning (Liu et al., 2020), WOODS (Katz-Samuels et al., 2022), and SCONE (Bai et al., 2023). Contrastly, our approach demonstrates strong performance on both OOD generalization and detection accuracy, which shows the effectiveness of our method for making use of the wild data.

E Results on Different Corruption Types

In this section, we provide additional ablation studies of the different covariate shifts. In Table 7, we evaluate our method under different common corruptions including Gaussian noise, shot noise, glass blur, and etc. To generate images with corruption, we follow the default setting and hyperparameters as in Hendrycks & Dietterich (2018). Our approach is robust under different covariate shifts and achieves strong OOD detection performance.

Table 6: Additional results. Comparison with competitive OOD detection and OOD generalization methods on CIFAR-10. For experiments using $\mathbb{P}_{\text{wild}}$, we use $\pi_s = 0.5$, $\pi_c = 0.1$. For each semantic OOD dataset, we create corresponding wild mixture distribution $\mathbb{P}_{\text{wild}} := (1 - \pi_s - \pi_c)\mathbb{P}_{\text{in}} + \pi_s \mathbb{P}_{\text{out}}^{\text{semantic}} + \pi_c \mathbb{P}_{\text{out}}^{\text{covariate}}$ for training. We report the average and std of our method based on **3 independent runs**. $\pm x$ denotes the rounded standard error.

Model	Places365 $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$				LSUN-R $\mathbb{P}_{\text{out}}^{\text{semantic}}$, CIFAR-10-C $\mathbb{P}_{\text{out}}^{\text{covariate}}$			
	OOD Acc.↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑
OOD detection								
MSP	75.05	94.84	57.40	84.49	75.05	94.84	52.15	91.37
ODIN	75.05	94.84	57.40	84.49	75.05	94.84	26.62	94.57
Energy	75.05	94.84	40.14	89.89	75.05	94.84	27.58	94.24
Mahalanobis	75.05	94.84	68.57	84.61	75.05	94.84	42.62	93.23
ViM	75.05	94.84	21.95	**95.48**	75.05	94.84	36.80	93.37
KNN	75.05	94.84	42.67	91.07	75.05	94.84	29.75	94.60
ASH	75.05	94.84	44.07	88.84	75.05	94.84	22.07	95.61
OOD generalization								
ERM	75.05	94.84	40.14	89.89	75.05	94.84	27.58	94.24
Mixup	79.17	93.30	58.24	75.70	79.17	93.30	32.73	88.86
IRM	77.92	90.85	53.79	88.15	77.92	90.85	34.50	94.54
VREx	76.90	91.35	56.13	87.45	76.90	91.35	44.20	92.55
EQRM	75.71	92.93	51.00	88.61	75.71	92.93	31.23	94.94
SharpDRO	79.03	**94.91**	34.64	91.96	79.03	94.91	13.27	97.44
Learning w. $\mathbb{P}_{wild}$								
OE	35.98	94.75	27.02	94.57	46.89	94.07	0.7	99.78
Energy (w/ outlier)	19.86	90.55	23.89	93.60	32.91	93.01	0.27	99.94
Woods	54.58	94.88	30.48	93.28	78.75	**95.01**	0.60	99.87
Scone	85.21	94.59	37.56	90.90	80.31	94.97	0.87	99.79
Ours	$\mathbf{89.16}_{\pm 0.01}$	$94.51_{\pm 0.11}$	$\mathbf{15.70}_{\pm 0.02}$	$94.68_{\pm 0.03}$	$\mathbf{91.08}_{\pm 0.01}$	$94.41_{\pm 0.00}$	$\mathbf{0.07}_{\pm 0.00}$	$\mathbf{99.98}_{\pm 0.00}$

Table 7: Ablation study on different corruption types for covariate OOD data. The budget is 500 for active learning. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Corruption types for covariate OOD data	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑
Gaussian noise	90.31	94.33	4.91	98.28
Defocus blur	94.54	94.68	1.38	99.56
Frosted glass blur	82.22	94.45	8.35	96.87
Impulse noise	91.82	94.38	3.61	98.88
Shot noise	91.98	94.62	3.79	98.82
Snow	92.51	94.45	2.89	99.04
Zoom blur	92.31	94.65	3.31	98.71
Brightness	94.59	94.53	1.56	99.48
Elastic transform	91.12	94.35	2.70	99.09
Contrast	94.15	94.64	1.56	99.50
Fog	94.34	94.67	1.26	99.53
Forst	92.42	94.49	3.19	98.79
Gaussian blur	94.40	94.69	1.14	99.58
Jpeg compression	89.75	94.52	3.37	98.95
Motion blur	92.44	94.38	2.89	99.14
Pixelate	93.08	94.42	2.28	99.24
Saturate	93.14	94.43	2.10	99.34
Spatter	93.66	94.60	1.98	99.35
Speckle noise	92.19	94.37	3.79	98.55

F Ablations on Mixing Rate λ

In the mixed sampling strategy, samples are selected using a combination of the top-k and near-boundary sampling methods. We introduce a mixing rate λ to determine the composition of the selected samples, where $0 \leq \lambda \leq 1$. Specifically, k samples are chosen from $\mathcal{S}_{\text{wild}}$. This set consists of k_1 samples from the top-k method and k_2 samples from the near-boundary sampling, where $k = k_1 + k_2$ and $\lambda = \frac{k_1}{k_1+k_2}$. In Table 8, we perform ablation on how λ affects the performance. We observe that a higher λ generally results

Table 8: Ablation results on different mixing rate λ. The total labeling budget is $k = 500$. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Mixing rate	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑
λ=0.1	87.93	94.90	16.08	92.17
λ=0.3	88.76	94.86	13.05	94.35
λ=0.5	88.08	95.02	12.98	94.41
λ=0.7	89.73	94.76	9.36	96.68
λ=0.9	89.80	94.70	8.05	97.91

in stronger performance in both OOD generalization and OOD detection. The observation aligns with our expectations and is supported by the detailed quantitative analysis presented in Section 4.3.

G Hyperparameter Analysis

In Table 9, we present the performance of OOD generalization and detection by varying hyperparameter α, which balances the weight between two loss terms. We observe that the generalization performance remains competitive and insensitive across a wide range of α values. Additionally, our method demonstrates enhanced OOD detection performance when a relatively larger value of α is employed in this scenario.

Table 9: Ablation study on the effect of loss weight α. The sampling strategy is top-k sampling, with a budget of 1000. We train on CIFAR-10 as ID, using wild data with $\pi_c = 0.5$ (CIFAR-10-C) and $\pi_s = 0.1$ (Texture).

Balancing weights	**OOD Acc.**↑	**ID Acc.**↑	**FPR**↓	**AUROC**↑
α=1.0	90.76	94.49	5.29	98.33
α=3.0	90.61	94.43	5.35	98.19
α=5.0	90.52	94.41	5.29	98.16
α=7.0	90.51	94.35	5.41	98.15
α=9.0	90.41	94.33	5.11	98.24
α=10.0	90.31	94.33	4.91	98.28

H Notations, Definitions, and Assumptions

Here we summarize important notations in Table 10, restate necessary definitions and assumptions in Sections H.2 and H.3.

H.1 Notations

Please see Table 10 for detailed notations.

Table 10: Main notations and their descriptions.

Notation	Description
	Spaces
$\mathcal{X}$, $\mathcal{Y}$	the input space and the label space.
$\mathcal{W}$, Θ	the hypothesis spaces
	Distributions
$\mathbb{P}_{\text{wild}}$, $\mathbb{P}_{\text{in}}$	data distribution for wild data, labeled ID data.
$\mathbb{P}_{\text{out}}^{\text{covariate}}$	data distribution for covariate-shifted OOD data.
$\mathbb{P}_{\text{out}}^{\text{semantic}}$	data distribution for semantic-shifted OOD data.
$\mathbb{P}_{\mathcal{X}\mathcal{Y}}$	the joint data distribution for ID data.
	Data and Models
$\mathbf{w}$, $\mathbf{x}$, $\mathbf{v}$	weight/input/the top-1 right singular vector of $\mathbf{G}$
$\widehat{\nabla}$, τ	the average gradients on labeled ID data, uncertainty score
$\mathcal{S}^{\text{in}}$, $\mathcal{S}_{\text{wild}}$	labeled ID data and unlabeled wild data
$\mathcal{S}_{\text{selected}}$	selected data
$\mathcal{S}^{\text{s}}_{\text{selected}}$, $\mathcal{S}^{\text{c}}_{\text{selected}}$	semantic and covariate OOD in the selected data $\mathcal{S}_{\text{selected}}$
$f_{\mathbf{w}}$ and $g_{\boldsymbol{\theta}}$	predictor on labeled in-distribution and binary predictor for OOD detection
y	label for ID classification
$\widehat{y}_{\mathbf{x}}$	Predicted one-hot label for input $\mathbf{x}$
n, m, k	size of $\mathcal{S}^{\text{in}}$, size of $\mathcal{S}_{\text{wild}}$, labeling budget
	Distances
$d_{\mathcal{W}\triangle\mathcal{W}}(\cdot,\cdot)$	$\mathcal{W}\triangle\mathcal{W}$ distance.
r_1 and r_2	the radius of the hypothesis spaces $\mathcal{W}$ and Θ, respectively
$\|\cdot\|_2$	ℓ_2 norm
	Loss, Risk and Predictor
$\ell(\cdot,\cdot)$	ID loss function
$R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{s}}_{\text{selected}}}(g_{\boldsymbol{\theta}})$	the overall empirical risk that classifies ID and detects semantic OOD
$R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})$	the overall empirical risk that classifies covariate OOD and ID
$R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}})$	the empirical risk w.r.t. predictor $f_{\mathbf{w}}$ over data $\mathcal{S}^{\text{in}}$
$R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})$	the empirical risk w.r.t. predictor $f_{\mathbf{w}}$ over covariate OOD $\mathcal{S}^{\text{c}}_{\text{selected}}$
ID-Acc	in-distribution accuracy.
OOD-Acc	out-of-distribution accuracy.
FPR	OOD detection performance.
	Additional Notations in Theory
ω_{in}, ω_{c}	the weight coefficients for ID empirical risk, and covariate OOD empirical risk.
$M = \beta_1 r_1^2 + b_1 r_1 + B_1$	the upper bound of loss $\ell(\mathbf{h}_{\mathbf{w}}(\mathbf{x}), y)$, see Proposition 3
d	VC dimension of the hypothesis space $\mathcal{W}$

H.2 Definitions

Definition 2 (β-smooth)**.** *We say a loss function $\ell(f_{\mathbf{w}}(\mathbf{x}), y)$ (defined over $\mathcal{X} \times \mathcal{Y}$) is β-smooth, if for any $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$,*

$$\left\|\nabla \ell(f_{\mathbf{w}}(\mathbf{x}), y) - \nabla \ell(f_{\mathbf{w}'}(\mathbf{x}), y)\right\|_2 \leq \beta \|\mathbf{w} - \mathbf{w}'\|_2$$

Definition 3 ($\mathcal{W}\triangle\mathcal{W}$-distance (Ben-David et al., 2010))**.** *For two distribution $\mathbb{P}_1$ and $\mathbb{P}_2$ over a domain $\mathcal{X}$ and a hypothesis class $\mathcal{W}$, the $\mathcal{W}\triangle\mathcal{W}$-distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ w.r.t. $\mathcal{W}$ is defined as*

$$d_{\mathcal{W}\triangle\mathcal{W}}(\mathbb{P}_1,\mathbb{P}_2) = \sup_{\mathbf{w},\mathbf{w}'\in\mathcal{W}} \left|\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_1}[f_{\mathbf{w}}(\mathbf{x}) \neq f_{\mathbf{w}'}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_2}[f_{\mathbf{w}}(\mathbf{x}) \neq f_{\mathbf{w}'}(\mathbf{x})]\right| \tag{9}$$

Definition 4 (Gradient-based Distribution Discrepancy)**.** *Given distributions $\mathbb{P}$ and $\mathbb{Q}$ defined over $\mathcal{X}$, the Gradient-based Distribution Discrepancy w.r.t. predictor $f_{\mathbf{w}}$ and loss ℓ is*

$$d^{\ell}_{\mathbf{w}}(\mathbb{P},\mathbb{Q}) = \left\|\nabla R_{\mathbb{P}}(f_{\mathbf{w}},\widehat{f}) - \nabla R_{\mathbb{Q}}(f_{\mathbf{w}},\widehat{f})\right\|_2, \tag{10}$$

where $\widehat{f}$ is a classifier which returns the closest one-hot vector of $f_{\mathbf{w}}$, $R_{\mathbb{P}}(f_{\mathbf{w}},\widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}}\ell(f_{\mathbf{w}},\widehat{f})$ and $R_{\mathbb{Q}}(f_{\mathbf{w}},\widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathbb{Q}}\ell(f_{\mathbf{w}},\widehat{f})$.

H.3 Assumptions

Assumption 1.

- *The parameter space $\mathcal{W} \subset B(\mathbf{w}_0, r_1) \subset \mathbb{R}^d$ (ℓ_2 ball of radius r_1 around $\mathbf{w}_0$);*
- *$\ell(f_{\mathbf{w}}(\mathbf{x}), y) \geq 0$ and $\ell(f_{\mathbf{w}}(\mathbf{x}), y)$ is β_1-smooth where $\ell(\cdot,\cdot)$ is the ID loss function;*
- $\sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \|\nabla\ell(f_{\mathbf{w}_0}(\mathbf{x}), y)\|_2 = b_1$, $\sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \ell(f_{\mathbf{w}_0}(\mathbf{x}), y) = B_1$.

Remark 1. For neural networks with smooth activation functions and softmax output function, we can check that the norm of the second derivative of the loss functions (cross-entropy loss and sigmoid loss) is bounded given the bounded parameter space, which implies that the β-smoothness of the loss functions can hold true. Therefore, our assumptions are reasonable in practice.

I Main Theorem

In this section, we provide a detailed and formal version of our main theorems with a complete description of the constant terms and other additional details that are omitted in the main paper.

Theorem 2. *Let $\mathcal{W}$ be a hypothesis space with a VC-dimension of d. Denote the datasets $\mathcal{S}^{in}$ and $\mathcal{S}^{c}_{selected}$ as the labeled in-distribution and the selected covariate OOD data by human feedback, and their sizes are n and m_c, respectively. If $\widehat{\mathbf{w}} \in \mathcal{W}$ minimizes the empirical risk $R_{\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}})$ of the multi-class classifier for classifying the covariate OOD, and $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}})$, then for any $\delta \in (0,1)$, with probability of at least $1-\delta$, we have*

$$R_{\mathbb{P}^{covariate}_{out}}(f_{\widehat{\mathbf{w}}}) \leq R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}^*}) + 2\omega_{in} \sup_{\mathbf{w}\in\mathcal{W}} d^{\ell}_{\mathbf{w}}(\mathcal{S}^{in}, \mathcal{S}^{c}_{selected}) + 2\omega_{in}(2\sqrt{\frac{2d\log(2m_c) + \log\frac{2}{\delta}}{m_c}} + \gamma) + 2\zeta,$$

where $\zeta = \sqrt{(\frac{\omega_{in}^2}{n} + \frac{\omega_c^2}{m_c})(\frac{d\log(2n+2m_c) - \log(\delta)}{2})} + \omega_{in}M$ and $\gamma = \min_{\mathbf{w}\in\mathcal{W}}\{R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}}) + R_{\mathbb{P}_{in}}(f_{\mathbf{w}})\}$. M is the upper bound of the loss function for the multi-class classifier

$$\sup_{\mathbf{w}\in\mathcal{W}} \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \ell(f_{\mathbf{w}}(\mathbf{x}), y) \leq M. \tag{11}$$

ω_{in}, ω_c are two weight coefficients such that

$$\underbrace{R_{\mathbb{P}_{in},\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}})}_{Multi\text{-}class\ classifier} = \omega_{in} R_{\mathbb{P}_{in}}(f_{\mathbf{w}}) + \omega_c R_{\mathbb{P}^{covariate}_{out}}(f_{\mathbf{w}}), \tag{12}$$

and $d^{\ell}_{\mathbf{w}}(\mathcal{S}^{in}, \mathcal{S}^{c}_{selected})$ is calculated as follows:

$$d^{\ell}_{\mathbf{w}}(\mathcal{S}^{in}, \mathcal{S}^{c}_{selected}) = \|\nabla R_{\mathcal{S}^{in}}(f_{\mathbf{w}}, \widehat{f}) - \nabla R_{\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}}, \widehat{f})\|_2,$$

where $\widehat{f}$ is a classifier which returns the closest one-hot vector of $f_{\mathbf{w}}$, i.e., $R_{\mathcal{S}^{in}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{in}}\ell(f_{\mathbf{w}}, \widehat{f})$ and $R_{\mathcal{S}^{c}_{selected}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{c}_{selected}}\ell(f_{\mathbf{w}}, \widehat{f})$. Note that our theoretical analysis primarily focuses on the OOD generalization error of a specific set of covariate data, which is associated with the set $\mathcal{S}^{c}selected$. For simplicity, we will continue to use the notation $\mathbb{P}^{covariate}_{out}$.

J Proof of the Main Theorem

In this section, we present the proof of our main Theorem 2. Before we dive into the proof details, we first clarify the analysis framework and the analysis target in our proof techniques.

Specifically, we consider the empirical error of the robust classification of samples from $\mathcal{S}^{\text{in}}$ and covariate OOD $\mathcal{S}^{\text{c}}_{\text{selected}}$ as the following weighted combination:

$$\underbrace{R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})}_{\text{Multi-class classifier}} = \omega_{\text{in}} R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}) + \omega_{\text{c}} R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}}). \tag{13}$$

Let $R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}})$ represents the error of $f_{\mathbf{w}}$ on the in distribution (ID) data $\mathbb{P}_{\text{in}}$, and $R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})$ denotes the error of $f_{\mathbf{w}}$ on the covariate OOD data $\mathbb{P}^{\text{covariate}}_{\text{out}}$. ω_{in} and ω_{c} denote the weight coefficients. Similarly, we can define the true risk over the data distributions in the same way:

$$\underbrace{R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})}_{\text{Multi-class classifier}} = \omega_{\text{in}} R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) + \omega_{\text{c}} R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}). \tag{14}$$

Step 1. First, we prove that for any $\delta \in (0,1)$ and $\mathbf{w} \in \mathcal{W}$, with probability of at least $1-\delta$, we have

$$P[|R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}) - R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})| \geq R] \leq \sqrt{(\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}}) - \log(\delta)}{2})}, \tag{15}$$

where n, m_{c} are the sizes of datasets $\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}$.

We first apply Theorem 3.2 of (Kifer et al., 2004) as restated in Lemma 8 to get the following equation,

$$P[|R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}) - R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})| \geq R] \leq (2n+2m_{\text{c}})^d \exp(\frac{-2R^2}{\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}}}),$$

where d is the VC dimension of the hypothesis space $\mathcal{W}$. Given $\delta \in (0,1)$, we set the upper bound of the inequality to δ, and solve for R:

$$\delta = (2n+2m_{\text{c}})^d \exp(\frac{-2R^2}{\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}}}).$$

We rewrite the inequality as

$$\frac{\delta}{(2n+2m_{\text{c}})^d} = e^{-2R^2/(\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}})},$$

taking the logarithm of both sides, we get

$$\log \frac{\delta}{(2n+2m_{\text{c}})^d} = -2R^2/(\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}}).$$

Rearranging the equation, we then get

$$R^2 = (\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}}) - \log(\delta)}{2}).$$

Therefore, with the probability of at least $1-\delta$, we have

$$|R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}) - R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})| \leq \sqrt{(\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}}) - \log(\delta)}{2})}. \tag{16}$$

Step 2. Based on Equation 16, we now prove Theorem 2. For the true error of hypothesis $\widehat{\mathbf{w}}$ on the covariate OOD data $R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\widehat{\mathbf{w}}})$, applying Lemma 7, Equation 16, and suppose $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})$, we get

$$
\begin{aligned}
R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\widehat{\mathbf{w}}}) &\leq R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\widehat{\mathbf{w}}}) + \omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) \\
&\leq R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\widehat{\mathbf{w}}}) + \sqrt{(\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}})-\log(\delta)}{2})} \\
&+ \omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) \\
&\leq R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}^*}) + \sqrt{(\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}})-\log(\delta)}{2})} \\
&+ \omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) \\
&\leq R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}^*}) + 2\sqrt{(\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}})-\log(\delta)}{2})} \\
&+ \omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) \\
&\leq R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}^*}) + 2\sqrt{(\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}})-\log(\delta)}{2})} \\
&+ 2\omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) \\
&= R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}^*}) + 2\omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) + 2\zeta_1,
\end{aligned}
$$

with probability of at least $1-\delta$, where $\zeta_1 = \sqrt{(\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}})(\frac{d\log(2n+2m_{\text{c}})-\log(\delta)}{2})}$ and $\gamma = \min_{h\in\mathcal{W}}\{R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}) + R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}})\}$

Step 3. In this step, we aim to obtain the upper bound of the term $d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}})$. To begin with, recall we have the following definition:

$$
d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) = \sup_{\mathbf{w},\mathbf{w}'\in\mathcal{W}}\left|\mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{\text{in}}}[f_{\mathbf{w}}(\mathbf{x})\neq f_{\mathbf{w}'}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{\text{c}}_{\text{selected}}}[f_{\mathbf{w}}(\mathbf{x})\neq f_{\mathbf{w}'}(\mathbf{x})]\right|. \tag{17}
$$

Therefore, it is easy to check that

$$
\begin{aligned}
d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) &= \sup_{\mathbf{w}\in\mathcal{W}}\big|R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}) - \nabla R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}},\widehat{f}) + \nabla R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}},\widehat{f}) - R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}}) \\
&\qquad - \nabla R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}},\widehat{f}) + \nabla R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}},\widehat{f})\big| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}}\big|R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}) - R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})\big| + 2\sup_{\mathbf{w}\in\mathcal{W}}\big|\nabla R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}},\widehat{f}) - \nabla R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}},\widehat{f})\big| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}}\big|R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}})\big| + \sup_{\mathbf{w}\in\mathcal{W}}\big|R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})\big| + 2\sup_{\mathbf{w}\in\mathcal{W}} d^{\ell}_{\mathbf{w}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) \\
&\leq 2\sup_{\mathbf{w}\in\mathcal{W}} d^{\ell}_{\mathbf{w}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 2M,
\end{aligned} \tag{18}
$$

where $\widehat{f}$ is a classifier which returns the closest one-hot vector of $f_{\mathbf{w}}$, $R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{\text{in}}}\ell(f_{\mathbf{w}}, \widehat{f})$ and $R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}}, \widehat{f}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{S}^{\text{c}}_{\text{selected}}}\ell(f_{\mathbf{w}}, \widehat{f})$.

The last inequality holds because of Proposition 3 and the definition of the Gradient-based Distribution Discrepancy in Definition 4. Therefore, we can prove that:

$$R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\widehat{\mathbf{w}}}) \leq R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}^*}) + 2\omega_{\text{in}} \sup_{\mathbf{w}\in\mathcal{W}} d^{\ell}_{\mathbf{w}}(\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}) + 2\omega_{\text{in}}(4\sqrt{\frac{2d\log(2m_{\text{c}}) + \log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma) + 2\zeta_1 + 2\omega_{\text{in}}M. \tag{19}$$

K Necessary Lemmas, and Propositions

K.1 Boundedness

Proposition 3. If Assumption 1 holds,

$$\sup_{\mathbf{w}\in\mathcal{W}} \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \|\nabla\ell(f_{\mathbf{w}}(\mathbf{x}),y)\|_2 \leq \beta_1 r_1 + b_1 = \sqrt{M'/2},$$

$$\sup_{\mathbf{w}\in\mathcal{W}} \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \ell(f_{\mathbf{w}}(\mathbf{x}),y) \leq \beta_1 r_1^2 + b_1 r_1 + B_1 = M,$$

Proof. One can prove this by *Mean Value Theorem of Integrals* easily. □

Proposition 4. If Assumption 1 holds, for any $\mathbf{w}\in\mathcal{W}$,

$$\left\|\nabla\ell(f_{\mathbf{w}}(\mathbf{x}),y)\right\|_2^2 \leq 2\beta_1\ell(f_{\mathbf{w}}(\mathbf{x}),y).$$

Proof. The details of the self-bounding property can be found in Appendix B of (Lei & Ying, 2021). □

Proposition 5. If Assumption 1 holds, for any labeled data $\mathcal{S}$ and distribution $\mathbb{P}$,

$$\left\|\nabla R_{\mathcal{S}}(f_{\mathbf{w}})\right\|_2^2 \leq 2\beta_1 R_{\mathcal{S}}(f_{\mathbf{w}}), \quad \forall\mathbf{w}\in\mathcal{W},$$

$$\left\|\nabla R_{\mathbb{P}}(f_{\mathbf{w}})\right\|_2^2 \leq 2\beta_1 R_{\mathbb{P}}(f_{\mathbf{w}}), \quad \forall\mathbf{w}\in\mathcal{W}.$$

Proof. Jensen's inequality implies that $R_{\mathcal{S}}(f_{\mathbf{w}})$ and $R_{\mathbb{P}}(f_{\mathbf{w}})$ are β_1-smooth. Then Proposition 4 implies the results. □

K.2 Necessary Lemmas for Theorem 2

Lemma 6 (Theorem 3.4 in Kifer et al. (2004)). Let $\mathcal{A}$ be a collection of subsets of some domain measure space, and assume that the VC-dimension is some finite d. Let P_1 and P_2 be probability distributions over that domain and S_1, S_2 finite samples of sizes m_1, m_2 drawn according to P_1, P_2 with certain selection criteria respectively. Then

$$P_{m1+m2}[|\phi_{\mathcal{A}}(S_1,S_2) - \phi_{\mathcal{A}}(P_1,P_2)| > R] \leq (2m_1)^d e^{-m_1R^2/16} + (2m_2)^d e^{-m_2R^2/16},$$

where P_{m1+m2} is the $m1+m2$'th power of P, the probability that P induces over the choice of samples.

This theorem bounds the probability for the relativized discrepancy, and will help bounds the quantified distribution shifts between domains in our Theorem 2.

Lemma 7. Let $\mathcal{W}$ be a hypothesis space with a VC-dimension of d. Denote the datasets $\mathcal{S}^{\text{in}}$ and $\mathcal{S}^{\text{c}}_{\text{selected}}$ as the labeled in-distribution and the selected covariate OOD data, and their sizes are n and m_{c}, respectively. Then for any $\delta\in(0,1)$, for every $\mathbf{w}\in\mathcal{W}$ minimizing $R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})$ on datasets $\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}$, we have

$$\|R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}) - R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})\| \leq \omega_{\text{in}}(\frac{1}{2}d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}) + 4\sqrt{\frac{2d\log(2m_{\text{c}})+\log\frac{2}{\delta}}{m_{\text{c}}}} + \gamma), \tag{20}$$

where $\gamma = \min_{\mathbf{w}\in\mathcal{W}}\{R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) + R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})\}$. $d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}})$ is defined according to Definition 3.

Proof. First, we prove that given datasets $\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}$ from two distributions $\mathbb{P}_{\text{in}}$ and $\mathbb{P}^{\text{covariate}}_{\text{out}}$, we have

$$d_{\mathcal{W}\triangle\mathcal{W}}(\mathbb{P}_{\text{in}}, \mathbb{P}^{\text{covariate}}_{\text{out}}) \leq d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}) + 4\sqrt{\frac{2d\log(2m_{\text{c}}) + \log\frac{2}{\delta}}{m_{\text{c}}}}. \tag{21}$$

We start with Theorem 3.4 in Kifer et al. (2004), which is restated in Lemma 6:

$$P_{n+m_{\text{c}}}[|\phi_{\mathcal{A}}(\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}) - \phi_{\mathcal{A}}(\mathbb{P}_{\text{in}}, \mathbb{P}^{\text{covariate}}_{\text{out}})| > R] \leq (2n)^d e^{-nR^2/16} + (2m_{\text{c}})^d e^{-m_{\text{c}}R^2/16}. \tag{22}$$

In this equation, d is the VC-dimension of a collection of subsets of some domain measure space $\mathcal{A}$, while in our case, d is the VC-dimension of hypothesis space $\mathcal{W}$. Following Ben-David et al. (2010), the VC-dimension of $\mathcal{W}\triangle\mathcal{W}$ is at most twice the VC-dimension of $\mathcal{W}$, and the VC-dimension of our domain measure space is thus $2d$.

Given $\delta \in (0, 1)$, we can set the upper bound of the inequality to δ, and solve for R:

$$\delta = (2n)^{2d} \cdot e^{-nR^2/16} + (2m_{\text{c}})^{2d} \cdot e^{-m_{\text{c}}R^2/16}. \tag{23}$$

Let $n = m_{\text{c}}$, we can rewrite the inequality as:

$$\frac{\delta}{(2m_{\text{c}})^{2d}} = e^{-n\epsilon^2/16} + e^{-m_{\text{c}}\epsilon^2/16}, \tag{24}$$

taking the logarithm of both sides, we get

$$\log\frac{\delta}{(2m_{\text{c}})^{2d}} = -n\frac{\epsilon^2}{16} + \log\left(1 + e^{-(n-m_{\text{c}})\frac{\epsilon^2}{16}}\right), \tag{25}$$

rearranging the equation and defining $a = \frac{R^2}{16}$, we then get

$$\log\frac{\delta}{(2m_{\text{c}})^{2d}} = -m_{\text{c}}a + \log 2, \tag{26}$$

which implies

$$m_{\text{c}}a + \log(\delta/2) = 2d\log(2m_{\text{c}}). \tag{27}$$

Therefore, we have

$$R = 4\sqrt{a} = 4\sqrt{\frac{2d\log(2m_{\text{c}}) + \log\frac{2}{\delta}}{m_{\text{c}}}}. \tag{28}$$

With probability of at least $1 - \delta$, we have

$$|\phi_{\mathcal{A}}(\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}) - \phi_{\mathcal{A}}(\mathbb{P}_{\text{in}}, \mathbb{P}^{\text{covariate}}_{\text{out}})| \leq 4\sqrt{\frac{2d\log(2m_{\text{c}}) + \log\frac{2}{\delta}}{m_{\text{c}}}}; \tag{29}$$

therefore,

$$d_{\mathcal{W}\triangle\mathcal{W}}(\mathbb{P}_{\text{in}}, \mathbb{P}^{\text{covariate}}_{\text{out}}) \leq d_{\mathcal{W}\triangle\mathcal{W}}(\mathcal{S}^{\text{in}}, \mathcal{S}^{\text{c}}_{\text{selected}}) + 4\sqrt{\frac{2d\log(2m_{\text{c}}) + \log\frac{2}{\delta}}{m_{\text{c}}}}. \tag{30}$$

Now in order to prove Lemma 7, we can use triangle inequality for classification error in the derivation.

For the true risk of hypothesis $f_{\mathbf{w}}$ on the covariate OOD data $R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})$, given the definition of $R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})$,

$$\begin{aligned}
\|R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})\| &= |\omega_{\text{in}} R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) + \omega_{\text{c}} R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})| \\
&\leq \omega_{\text{in}} |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})| \\
&\leq \omega_{\text{in}} (|R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) - R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})| + |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})| \\
&\qquad + |R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})|) \\
&\leq \omega_{\text{in}} (R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}^*}) + |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})| + R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}^*})) \\
&\leq \omega_{\text{in}} (\gamma + |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})|),
\end{aligned}$$

where $\gamma = \min_{h \in \mathcal{W}} \{R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}) + R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})\}$ and $f_{\mathbf{w}^*}$ is classifier that are parameterized with the optimal hypothesis $f_{\mathbf{w}^*}$ on $\mathbb{P}_{\text{in}}$. And we also have

$$R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[|f_{\mathbf{w}}(\mathbf{x}) - f_{\mathbf{w}^*}(\mathbf{x})|]. \tag{31}$$

By the definition of $\mathcal{W} \triangle \mathcal{W}$-distance and our proved Equation 30,

$$\begin{aligned}
\|R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})\| &\leq \sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}'}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}'})| \\
&= \sup_{\mathbf{w}, \mathbf{w}' \in \mathcal{W}} P_{\mathbf{x} \sim \mathbb{P}_{\text{in}}}[f_{\mathbf{w}}(\mathbf{x}) \neq f_{\mathbf{w}^*}(\mathbf{x})] + P_{\mathbf{x} \sim \mathbb{P}_{\text{out}}^{\text{covariate}}}[f_{\mathbf{w}}(\mathbf{x}) \neq f_{\mathbf{w}^*}(\mathbf{x})] \\
&= \frac{1}{2} d_{\mathcal{W} \triangle \mathcal{W}}(\mathbb{P}_{\text{in}}, \mathbb{P}_{\text{out}}^{\text{covariate}}) \\
&\leq \frac{1}{2} d_{\mathcal{W} \triangle \mathcal{W}}(\mathcal{S}^{\text{in}}, \mathcal{S}_{\text{selected}}^{\text{c}}) + 2\sqrt{\frac{2d \log(2m_{\text{c}}) + \log \frac{2}{\delta}}{m_{\text{c}}}}.
\end{aligned}$$

Therefore, we can get

$$\begin{aligned}
|R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})| &\leq \omega_{\text{in}} (\gamma + |R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*}) - R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}, f_{\mathbf{w}^*})|) \\
&\leq \omega_{\text{in}} (\gamma + \frac{1}{2} d_{\mathcal{W} \triangle \mathcal{W}}(\mathbb{P}_{\text{in}}, \mathbb{P}_{\text{out}}^{\text{covariate}})) \\
&\leq \omega_{\text{in}} (\frac{1}{2} d_{\mathcal{W} \triangle \mathcal{W}}(\mathcal{S}^{\text{in}}, \mathcal{S}_{\text{selected}}^{\text{c}}) + 2\sqrt{\frac{2d \log(2m_{\text{c}}) + \log \frac{2}{\delta}}{m_{\text{c}}}} + \gamma),
\end{aligned}$$

with probability of at least $1 - \delta$, where $\gamma = \min_{\mathbf{w} \in \mathcal{W}} \{R_{\mathbb{P}_{\text{in}}} + R_{\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}})\}$. This completes the proof.

□

Lemma 8. Under the same conditions as Lemma 7, if the empirical risk is denoted as $R_{\mathcal{S}^{\text{in}},\mathcal{S}_{\text{selected}}^{\text{c}}}(f_{\mathbf{w}})$ (as defined in Equation 13), then for any $\delta \in (0, 1)$ and $\mathbf{w} \in \mathcal{W}$, with the probability of at least $1 - \delta$, we have

$$P[|R_{\mathbb{P}_{\text{in}},\mathbb{P}_{\text{out}}^{\text{covariate}}}(f_{\mathbf{w}}) - R_{\mathcal{S}^{\text{in}},\mathcal{S}_{\text{selected}}^{\text{c}}}(f_{\mathbf{w}})| \geq R] \leq 2 \exp(\frac{-2R^2}{\frac{\omega_{\text{in}}^2}{n} + \frac{\omega_{\text{c}}^2}{m_{\text{c}}}}) \tag{32}$$

Proof. We apply Hoeffding's Inequality in our proof. Specifically, denote the true labeling function as $f_{\bar{\mathbf{w}}}$, we have that

$$\mathbb{P}(|[\sum_{i=0}^{n}\frac{\omega_{\text{in}}}{n}|f_{\mathbf{w}}(\mathbf{x}_i)-f_{\bar{\mathbf{w}}}(\mathbf{x}_i)|+\sum_{j=0}^{m_{\text{c}}}\frac{\omega_{\text{c}}}{m_{\text{c}}}|f_{\mathbf{w}}(\mathbf{x}_j)-f_{\bar{\mathbf{w}}}(\mathbf{x}_j)|]-[\omega_{\text{in}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{in}}}|f_{\mathbf{w}}(\mathbf{x})-f_{\bar{\mathbf{w}}}(\mathbf{x})|+\omega_{\text{c}}\mathbb{E}_{x\sim\mathbb{P}_{\text{c}}}|f_{\mathbf{w}}(\mathbf{x})-f_{\bar{\mathbf{w}}}(\mathbf{x})|]|\geq R)$$
$$\leq 2\exp(-\frac{2R^2}{\sum_{i=1}^{n}(b_i-a_i)^2}), \tag{33}$$

where $\frac{\omega_{\text{in}}}{n}|f_{\mathbf{w}}(\mathbf{x}_i)-f_{\bar{\mathbf{w}}}(\mathbf{x}_i)|\in[0,\frac{\omega_{\text{in}}}{n}]$ and $\frac{\omega_{\text{c}}}{m_{\text{c}}}|f_{\mathbf{w}}(\mathbf{x}_j)-f_{\bar{\mathbf{w}}}(\mathbf{x}_j)|\in[0,\frac{\omega_{\text{c}}}{m_{\text{c}}}]$.

Considering the weighted empirical error, we get

$$\begin{aligned}R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})&=\omega_{\text{in}}R_{\mathcal{S}^{\text{in}}}(f_{\mathbf{w}})+\omega_{\text{c}}R_{\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})\\&=\sum_{i=0}^{n}\frac{\omega_{\text{in}}}{n}|f_{\mathbf{w}}(\mathbf{x}_i)-f_{\bar{\mathbf{w}}}(\mathbf{x}_i)|+\sum_{j=0}^{m_c}\frac{\omega_{\text{c}}}{m_{\text{c}}}|f_{\mathbf{w}}(\mathbf{x}_j)-f_{\bar{\mathbf{w}}}(\mathbf{x}_j)|,\end{aligned}$$

which corresponds to the first part of Hoeffding's Inequality.

Due to the linearity of expectations, we can calculate the sum of expectations as

$$\omega_{\text{in}}\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{in}}}|f_{\mathbf{w}}(\mathbf{x})-f_{\bar{\mathbf{w}}}(\mathbf{x})|+\omega_{\text{c}}\mathbb{E}_{x\sim\mathbb{P}_{\text{c}}}|f_{\mathbf{w}}(\mathbf{x})-f_{\bar{\mathbf{w}}}(\mathbf{x})|=\omega_{\text{in}}R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}})+\omega_{\text{c}}R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})=R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}}),$$

which corresponds to the second part of Hoeffding's Inequality. Therefore, we can apply Hoeffding's Inequality as

$$P[|R_{\mathbb{P}_{\text{in}},\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})-R_{\mathcal{S}^{\text{in}},\mathcal{S}^{\text{c}}_{\text{selected}}}(f_{\mathbf{w}})|\geq R]\leq 2\exp(\frac{-2R^2}{\frac{\omega_{\text{in}}^2}{n}+\frac{\omega_{\text{c}}^2}{m_{\text{c}}}}).$$

□

L Verification of Main Theorem

Optimal Loss for Covariate OOD. For training the model, we utilized 50,000 covariate OOD data samples. The optimal loss for covariate OOD data, denoted as $R_{\mathbb{P}^{\text{covariate}}_{\text{out}}}(f_{\mathbf{w}})$, was evaluated using the CIFAR-10 versus CIFAR-10-C datasets (with Gaussian noise). The results indicated an optimal loss of 0.2383 on the test set, with a corresponding OOD test accuracy of 92.79%. This small optimal loss for covariate OOD data contributes to a tighter upper bound.

Optimal Loss for In-Distribution (ID) Data. The training involved 50,000 ID data samples to determine the optimal loss for ID data, represented as $R_{\mathbb{P}_{\text{in}}}(f_{\mathbf{w}})$. In the CIFAR-10 context, the optimal loss on the test set for ID data was recorded as 0.1792, while the corresponding ID accuracy on test data reached 95.13%. The minimal nature of the optimal loss for ID data is consistent with expectations and results in a tighter upper bound.

Gradient Discrepancy. The gradient discrepancy for the ID CIFAR-10, Covariate OOD CIFAR-10-C (Gaussian noise), and Semantic OOD Textures dataset was found to be 0.00035. This small gradient discrepancy suggests a tighter upper bound.

Dataset	Gradient Discrepancy↓	OOD Acc.↑
CIFAR-10-C (Gaussian noise)	0.00035	90.37
CIFAR-10-C (Shot noise)	0.00030	82.04
CIFAR-10-C (Glass blur)	0.00040	92.41

Table 11: Empirical verification of gradient discrepancy in Theorem 1.

Gradient Discrepancy Versus Covariate OOD Accuracy Across Different Datasets. Table 11 offers a comparative analysis, empirically validating the gradient discrepancy among various datasets. The results show a correlation between gradient discrepancy and OOD accuracy.

M Impact Statements and Limitations

Broader Impact. Our research aims to raise both research and societal awareness regarding the critical challenges posed by OOD detection and generalization in real-world contexts. On a practical level, our study has the potential to yield direct benefits and societal impacts by ensuring the safety and robustness of deploying classification models in dynamic environments. This is particularly valuable in scenarios where practitioners have access to unlabeled datasets and need to discern the most relevant portions for safety-critical applications, such as autonomous driving and healthcare data analysis. From a theoretical standpoint, our analysis contributes to a deeper understanding of leveraging unlabeled wild data by using gradient-based scoring for selecting the most informative samples for human feedback. In Appendix L, we properly verify the necessary conditions of our bound using real-world datasets. Hence, we believe our theoretical framework has a broad utility and significance.

Limitations. Our proposed algorithm aims to improve both out-of-distribution detection and generalization results by leveraging unlabeled data. It still requires a small amount of human annotations and an additional gradient-based scoring procedure for deployment in the wild. Therefore, extending our framework to further reduce the annotation and training costs is a promising next step.

N Software and Hardware

We run all experiments with Python 3.8.5 and PyTorch 1.13.1, using NVIDIA GeForce RTX 2080 Ti GPUs.