LOG: Active Model Adaptation for Label-Efficient OOD Generalization

Jie-Jing Shao, Lan-Zhe Guo, Xiao-Wen Yang, Yu-Feng Li*
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{shaojj, guolz, yangxw, liyf}@lamda.nju.edu.cn

Abstract

This work discusses how to achieve worst-case Out-Of-Distribution (OOD) generalization for a variety of distributions based on a relatively small labeling cost. The problem has broad applications, especially in non-i.i.d. open-world scenarios. Previous studies either rely on a large amount of labeling cost or lack of guarantees about the worst-case generalization. In this work, we show for the first time that active model adaptation could achieve both good performance and robustness based on the invariant risk minimization principle. We propose LOG, an interactive model adaptation framework, with two sub-modules: active sample selection and causal invariant learning. Specifically, we formulate the active selection as a mixture distribution separation problem and present an unbiased estimator, which could find the samples that violate the current invariant relationship, with a provable guarantee. The theoretical analysis supports that both sub-modules contribute to generalization. A large number of experimental results confirm the promising performance of the new algorithm.

1 Introduction

Machine learning models are typically trained and tested on the same data distribution. However, when these models are deployed in real task scenarios, they face much inapplicability, because the data distribution of the target task usually deviates from that of training. For example, the financial data prediction model based on local users is often inaccurate in predicting user behavior in other regions. Similar examples include self-driving [41], speech recognition [16], influenza detection [30], etc.

To this problem, the recent research on *OOD (Out-Of-Distribution) Generalization* [23, 7, 1, 13, 39, 19, 5] has given a series of technologies, trying to obtain models with robustness, which have certain worst-case generalization guarantees on a variety of unseen distributions. Although they do not access target data, they need a large number of high-quality source data to remove the source-specific spurious correlation and obtain the generalizable invariant relationship, i.e., sufficient multi-source labeled data and accurate source information. However, these two requirements are still difficult to satisfy in most practical tasks.

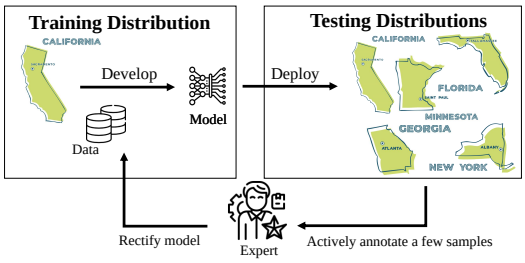


Figure 1: Example of model adaptation to a variety of distributions. We would like to annotate a few samples from target testing distributions to rectify the model in the source iteratively.

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

There are also some more traditional approaches to handle distribution shift, such as *Domain Adaptation* [25, 9, 20, 21]. They introduce unlabeled data from the target distribution, reducing the requirements on the source data relative to OOD generalization. However, they typically assume that target samples are from an isolated homogeneous distribution, which ignores the generalization in varied distributions. Recently, there are some works [10, 26, 22, 33] focusing on the adaptation to multiple latent domains. Nevertheless, the robustness, i.e., worst-case guarantees has not been addressed, especially without explicit domain labels.

To make a reasonable compromise between strongly labeled information dependence and robustness, *Active Model Adaptation* may be a feasible scheme. It gradually increases the labeling cost and actively annotates the data in the target task that is difficult to be robustly generalized. Such a scheme could effectively reduce the strong dependence on label information. It has shown promising performance when the target distribution is an isolated homogeneous distribution [38, 28, 6, 8].

In this paper, we would like to derive a benefit from *Active Model Adaptation* to address robustness at a small labeling cost. We first propose the invariant risk minimization principle for active model adaptation. That is, we could tend to both overall performance and worst-case guarantee via maximal invariant predictor. Based on this, we further present an interactive framework to achieve label-efficient OOD generalization (LOG), composed of an actively querying module and an invariant learning module. The key challenge is to find unlabeled samples where the current invariant model does not hold. Based on the structural causal model assumption, we formulate it as a mixture distribution separation problem and present an unbiased estimator to address it. In theory, our actively querying module could accurately find the samples which violate the current invariant relationship. In the experiments on a series of tasks, the promising performance of our LOG has been confirmed.

2 Related Work

OOD (Out-of-Distribution) Generalization works on learning models that generalize well on a variety of unseen distributions. There are two main branches: *Domain Generalization* [23, 7, 39] mostly focuses on computer vision problems as predictions are prone to be affected by a disturbance on images (e.g., style, background, etc). *Causal & Invariant Learning* starts from causal inference and explores causal variables to address generalization ability under covariate shift [1, 13]. Most of them rely on labeled data from multiple sources to remove the source-specific spurious correlation. Recently, some works [19, 5] attempt to mine the heterogeneity of an assembled source without explicit prior division. Nevertheless, they commonly assume source data is sufficient to learn the invariance and ignore the exploration of data from target distributions.

Domain Adaptation works on addressing the domain shift between the training source and the testing target, where no labeled data are available in the target domain [25, 9, 20, 21]. They typically assume the target samples are from a single homogeneous domain. The most related subtopic to us is *Latent Domain & Domain-Agnostic Adaptation* that focuses on the target distribution with multiple latent domains [26, 22, 33]. Although they propose methods with good overall performance in the presence of multiple latent distributions, the worst-case guarantee has not been addressed.

Active Model Adaptation works on active learning under distribution shift. Previous works have focused on two types of distribution shifts. 1) label space shift: adapt a pre-trained model to a task with different label spaces [11]; 2) domain shift: adapt a source model to a target domain [38, 28, 6, 8, 34]. To the best of our knowledge, all of them directly regard the target distribution as a homogeneous domain and ignore the generalization ability when massive distributions exist.

3 Problem and Analysis

Generally speaking, the above technologies are difficult to directly deal with the model generalization of a variety of distributions under a small labeling cost. To deal with such a challenge, in this paper, focusing on active model adaptation, we first analyze the theoretical basis of the problem. Based on this, we present a new active adaptation principle, invariant risk minimization. We further put forward the corresponding algorithm, and show its effect on generalization ability.

3.1 Problem Formulation

In the source training stage, the learner collects a sufficient labeled dataset $D_S = \{D^e\}_{e \in \mathcal{E}_S}$, which is a mixture of data $D^e \in \mathcal{X} \times \mathcal{Y}$ collected from the collection of training sources $e \in \mathcal{E}_S$. $\mathcal{X}$ and $\mathcal{Y}$ denote an input and output space, respectively. Source model $f_S : \mathcal{X} \to \mathcal{Y}$ is well-trained on the source dataset D_S with a small risk $R(f_S; D_S)$. When the model is deployed in open-world scenarios, it needs to adapt to the distributions $\mathcal{E}_T$, i.e. a collection of varying testing distributions. Following [26, 33], we formulate the target distribution $\mathcal{D}_T \in \mathcal{X} \times \mathcal{Y}$ as a mixture of base distributions: $\mathcal{D}_T = \sum_{e \in \mathcal{E}_T} \lambda^e \mathcal{D}^e$, where $\mathcal{D}^e$ is the distribution observed from the environment e and λ^e represents the corresponding proportion. Note that in reality, data are frequently assembled under implicit environment information e and λ^e, thus we do not depend on this information to develop algorithms. They are used only during evaluation. To fast adapt the source model to the target distribution, active learning has been introduced [38, 28, 8, 34]. Generally, their goal is:

Definition 3.1 (Performance Maximization)**.** The performance goal is to minimize the generalization risk of model f on the overall target distribution $\mathcal{D}_T$ based on some queried samples Q:

$$\min_{Q,\mathcal{A}} R(\mathcal{A}\left(f_S; \{D_S, Q\}\right); \mathcal{D}_T) \tag{1}$$

where $\mathcal{A}$ is a model adaptation algorithm to rectify f_S: $f \leftarrow \mathcal{A}\left(f_S; \{D_S, Q\}\right)$.

In addition to the optimization of ideal overall generalization, we also need to comprehensively consider the robustness, i.e., worst-case generalization on $\mathcal{E}_T$:

Definition 3.2 (Robustness Preservation)**.** The robustness means we could maintain low risk even in the worst distribution $\mathcal{D}^e$ across $\mathcal{E}_T$, i.e. worst-case guarantee on each base distribution $\mathcal{D}^e$.

$$\min_{Q,\mathcal{A}} \max_{e \in \mathcal{E}_T} R(\mathcal{A}\left(f_S; \{D_S, Q\}\right); \mathcal{D}^e) \tag{2}$$

Q and $\mathcal{A}$ are important to address active model adaptation. Previous works [11, 38, 28, 8] mainly focus on the Q to address performance maximization. They do not consider the choice of $\mathcal{A}$ (simply take it as standard supervised learning or semi-supervised domain adaptation) and the robustness.

3.2 Invariant Risk Minimization

Without any prior knowledge or structural assumptions, it is impossible to adapt the source model to target distributions, since one cannot characterize the shift between $\mathcal{E}_S$ and $\mathcal{E}_T$. Following the [1, 13, 2, 24, 37, 17], we consider the data generation via a structural causal model under the covariate shift assumption (i.e., the $P(y|x)$ is unchanged across the varying distributions):

Assumption 3.3. Consider a structural causal model [40] governing the random distribution $P(X, Y)$ and the learning goal of predicting Y from X. Then the set of all distributions $\mathcal{E}$ indexes all the interventional distribution $P^e(X^e, Y^e)$ obtainable by valid interventions e:

$$X^e = S(Z_s, Z_v^e), Y^e = f(Z_s) + \epsilon, \epsilon \perp X^e. \tag{3}$$

Z_s and Z_v^e represent the semantic variable and intervention variable, respectively. The feature X^e could be regrad as an observation S for Z_s in the intervention variable e, influenced by the Z_v^e. The label Y is caused by the semantic Z^e and an independent noise term ϵ. We assume the data generation process S is inevitable and there exists $\Phi(S(X^e)) = Z_s$ to recover the semantics for all Z_s and Z_v^e.

Remark 3.4. *Following the above assumption about structural causal model, the ideal Φ^* satisfies: 1) The margirial invariance $P^e(\Phi^*(X)) = P^{e'}(\Phi^*(X))$ and conditional invariance $P^e(Y|\Phi^*(X)) = P^{e'}(Y|\Phi^*(X))$ hold for any $e, e' \in \mathcal{E}$. 2) It is sufficient to predict the target using $\Phi^*(X)$ as the input: $Y = f(\Phi^*(X)) + \epsilon, \epsilon \perp X$, across varying $e \in \mathcal{E}$.*

To acquire such $\Phi^*(X)$, a branch of invariant risk minimization proposals [1, 13, 19, 5] finds the *Invariant Features* and the corresponding *Maximal Invariant Predictor*, defined as:

Definition 3.5 (Invariant Features)**.** The set of invariant features $\mathcal{I}$ with respect to $\mathcal{E}$ is defined as:

$$\mathcal{I}_{\mathcal{E}} = \{\Phi(X) : H[Y|\Phi(X)] = H[Y|\Phi(X), \mathcal{E}], \Phi(X) \perp \mathcal{E}\}$$

where $H[\cdot]$ is Shannon entropy of a random variable. The corresponding *Maxiaml Invariant Predictor* (MIP) of $\mathcal{I}_{\mathcal{E}}$ is defined as: $\Phi^* = \arg\max_{\Phi \in \mathcal{I}_{\mathcal{E}}} I(Y; \Phi)$ where $I(\cdot;\cdot)$ measures Shannon mutual information between two random variables.

Here we prove that we could achieve both performance and robustness based on the MIP Φ^*.

Theorem 3.6. *For a predictor $\Phi^*(X)$ in the collection $\mathcal{E}$, the solution $f^*(X) = P(Y|\Phi^*)$ could achieve both performance maximization $R(f^*; \mathcal{D}_T) \leq \min_f R(f; \mathcal{D}_T)$ and worst-case robustness $\max_{e\in\mathcal{E}} R(f^*; \mathcal{D}^e) \leq \min_f \max_{e\in\mathcal{E}} R(f; \mathcal{D}^e)$.*

Although the ideal MIP has shown a favorable theoretical guarantee, we find its generalization is heavily dependent on the heterogeneity of source distributions collection $\mathcal{E}_S$. We further analyze the connection between generalization and source heterogeneity and propose an active model adaptation framework to expand the heterogeneity and trend to the ideal MIP.

3.3 Active Heterogeneity Expansion

Given source collection $\mathcal{E}_S$ and the target collection $\mathcal{E}_T$, denote the corresponding invariant features set $\mathcal{I}_S$ and $\mathcal{I}_T$ respectively. For $\mathcal{E}_S \subseteq \mathcal{E}_T$, the corresponding invariant features set satisfies $\mathcal{I}_T \subseteq \mathcal{I}_S$ [19]. It indicates the generalization of $\mathcal{I}_S$ could be improved by extending the source $\mathcal{E}_S$ and promoting the $\mathcal{I}_S$ to ideal $\mathcal{I}_T$. Here, we present the generalization condition for Φ_S to an unseen distribution e' and then justify that the generalization of Φ_S is heavily dependent on the heterogeneity of $\mathcal{E}_S$.

Theorem 3.7. *For distribution $P^{e'}(X^{e'}, Y^{e'})$, if $\Phi(X^{e'}) = \Phi(X^e)_{e\in\mathcal{E}_S}$, $\mathcal{I}_S$ is equal to the invariance set constrained by $\mathcal{E}_S \cup \{e'\}$. The optimal source model f_S could generalize on the distribution $P^{e'}$. The generaliable distributions: $\mathcal{E}_G = \{e'|\mathcal{I}_S = \mathcal{I}_{\mathcal{E}_S\cup\{e'\}}\} = \{e'|\Phi(X^{e'}) = \Phi(X^e), e \in \mathcal{E}_S\}$.*

Particularly in the linear structural causal model, we further assume the semantics Z_s takes values in $\mathbb{R}^c$, intervention Z_v^e takes values in $\mathbb{R}^w$, and $S \in \mathbb{R}^{d\times(c+w)}$. Let $\Phi \in R^{d\times d}$, we have:

$$\mathcal{E}_G = \{e'|\mathbb{E}[X^{e'}] = \mathbb{E}[X^e] - x, e \in \mathcal{E}_S, x \in \ker(\Phi)\} \tag{4}$$

where $\dim(\ker(\Phi)) = \dim(\mathrm{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})) - 1$.

Remark 3.8. *Theorem 3.7 propose the generalization condition for any unseen e'. Compared with previous analysis [1, 27] in linear cases, the Equation 4 further indicates a quantitative dependence of generalization on source heterogeneity. In words, the freedom of $\mathcal{E}_G$ and the nullspace $\ker(\Phi)$ is limited by the maximal linearly independent system (heterogeneity) from source collection $\mathcal{E}_S$.*

Theorem 3.7 motives us to expand collection $\mathcal{E}_S$ via querying samples from a distribution $P^{e'}$ where $\Phi(X^{e'}) \neq \Phi(X^e)_{e\in\mathcal{E}_S}$. Formally, we have the following proposition:

Proposition 3.9 (Active Heterogeneity Expansion)**.** *To address performance maximization and robustness, we could guide the active model adaptation via the heterogeneity expansion:*

$$\mathcal{A}(f_S; Q) \to f^* = P(Y|\Phi_T^*), \tag{5}$$

promoting the Φ_S of f_S to ideal Φ_T. At each active data collection stage, we query and collect the samples from $\mathcal{E}_T \backslash \mathcal{E}_G$ to expand the data heterogeneity.

Theorem 3.10. *Under the linear structural causal model, when we collect a distribution e' where $\Phi(X^{e'}) \neq \Phi(X^e)_{e\in\mathcal{E}_S}$, and update $\Phi \to \Phi'$, $\mathrm{rank}(\Phi') = \mathrm{rank}(\Phi) - 1$ holds. The Φ will converge to the ideal Φ^* at most w iterations, where w is the freedom of the intervention variable $Z_v^e \in \mathbb{R}^w$.*

Theorem 3.10 indicates our proposed active heterogeneity expansion framework shares a favorable convergence rate. In words, each time we collect a distribution from $\mathcal{E}_T \backslash \mathcal{E}_G$, we can remove one degree of freedom in the space of the variant intervention factor. It is noteworthy that although the convergence analysis is based on the linear condition, our proposed framework is still applicable in non-linear cases where the condition $\mathcal{E}_G = \{e'|\Phi(X^{e'}) = \Phi(X^e), e \in \mathcal{E}_S\}$ in Theorem 3.7 holds.

4 Algorithm

Following the above analysis, we would like to query the samples from $\mathcal{E}_T \backslash \mathcal{E}_G$ to achieve the ideal $\mathcal{I}_T$. In this work, we propose our active model adaptation method LOG, with two interactive modules:

1) *Query Strategy* $\mathcal{M}_Q$: given the unlabeled data pool X_T and the current invariant relationship $\mathcal{I}$, actively query the representative samples Q from the un-generalizable distributions $\mathcal{E}_U = \mathcal{E}_T \backslash \mathcal{E}_G$.

2) *Model Adaptation* $\mathcal{M}_A$: given the newly queried samples Q, update the invariant relationship $\mathcal{I}$ via invariant learning.

The whole framework is iterative so that the mutual promotion between active exploration and invariance exploitation can be leveraged.

4.1 Query Strategy Module $\mathcal{M}_Q$

Following the problem formulation, we have the $X_T = \theta X_G + (1-\theta)X_U$, where $X_G = P(X|\mathcal{E}_G)$, $X_T = P(X|\mathcal{E}_T)$. The X_U reperesents the obversation from the un-generalizable $\mathcal{E}_U = \mathcal{E}_T \backslash \mathcal{E}_G$.

Based on Proposition 3.9, we can query few instances Q from X_U to promote $\Phi_{\{D_S,Q\}} \to \Phi_T^*$. This goal could be formulated as:

$$\mathcal{M}_A([X_s, y_s], [X_Q, y_Q]) \to \mathcal{M}_A([X_s, y_s], [X_U, y_U])$$

We first detect the samples from X_U and then find the representative instances X_Q for X_U to query. Based on the connection $\Phi(X_S) = \Phi(X_G)$, we could transform the detection of X_U as a mixture distribution separation problem. We denote $\mathbb{I}(x_i \sim X_G)$ to indicate if x is observed from the $\mathcal{E}_G$. The indicator is essentially a binary classifier $g \circ \Phi(x) : \mathcal{X} \to \{1, -1\}$. Given a sample x, when $g \circ \Phi(x) = 1$, it is observed from $\mathcal{E}_G$, and the current model could give a correct prediction, otherwise it will violate the current invariance Φ and is risky for the current model.

One of the main challenges to learning g is that we do not have information about the marginal distribution of unseen $\mathcal{E}_U$. We handle this problem by using the risk rewriting technique [12, 42] with the unlabeled data from $\mathcal{E}_T$.

$$P(\Phi(X_U)) = (P(\Phi(X_T)) - \theta P(\Phi(X_G)))/(1-\theta) \tag{6}$$

Proposition 4.1. *For all measurable function g, we could rewrite the original risk as:*

$$\begin{aligned} R(g) &= \theta E_{x\sim X_G}\left[\ell(g\circ\Phi(x), 1)\right] + (1-\theta)E_{x\sim X_U}\left[\ell(g\circ\Phi(x), -1)\right] \\ &= \theta E_{x\sim X_G}\left[\ell(g\circ\Phi(x), 1) - \ell(g\circ\Phi(x), -1)\right] + E_{x\sim X_T}[\ell(g\circ\Phi(x), -1)] \end{aligned}$$

Since the risk equals the ideal risk, its empirical estimator $\hat{R}(g)$ is unbiased over the target distribution. We can thus perform the standard empircal risk minimization. Particularly when the binary loss satisfying $\ell(z) - \ell(-z) = -z$, for all $z \in \mathbb{R}$, the empirical risk $\hat{R}(g)$ is convex w.r.t. g [12] [2]. It leads to a convex optimization problem, for which the globally optimal solution can be obtained.

Here we further analyze the estimation error of indictor g based on [12]. Formally, the empirical estimator $\hat{g}$ has the estimation error bound:

Theorem 4.2 (Estimation Error Bound)**.** *Suppose that* $\inf_{g\in\mathcal{G}} R(g) \geq \alpha > 0$ *and* $\mathcal{G}$ *is closed under negation, i.e.* $g \in \mathcal{G}$ *iff* $-g \in \mathcal{G}$*. Then for any* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$R(\hat{g}) - \inf_{g\in\mathcal{G}} R(g) \leq \mathcal{O}_p\left(\frac{1}{\sqrt{N_S}} + \frac{1}{\sqrt{N_T}}\right). \tag{7}$$

For a better presentation, we use the $\mathcal{O}_p$*-notation to keep the dependence on* N_S *and* N_T *only.*

Theorem 4.2 shows that the estimation error of the estimated $\hat{g}$ decreases with a growing number of source data N_S and unlabeled target data N_T. In our problem, we have the plentiful source data D_S from $\mathcal{E}_S$ and *unlabeled data* X_T from $\mathcal{E}_T$, which helps to estimate $\hat{g}$ accurately.

Notice that the implementation of our algorithm requires the knowledge of the mixture proportion θ, where plenty of works [3, 31, 29] have been explored to estimate θ from the unlabeled data. We adopt the mixture proportion estimation method of [29] in our implementation and omit the details here. All of our empirical studies are conducted by the estimated mixture proportion.

Based on the distribution inference module g, we could get the probability $p_u(x) = p(-1|x; g\circ\Phi)$ that x belongs to $\mathcal{E}_U$. Here we further consider to obtain the limited subset Q covering the empirical set $X_U = p_u(x) \cdot X_T$. Specifically, we follow the core-set objective proposed in [32]:

$$\min_{Q} \max_{x\in X_U} \min_{x'\in Q} \|x - x'\|_2 \leq \delta_Q \tag{8}$$

[2] Many popular loss functions satisfy the condition, such as logistic loss, square loss and double hinge loss.

Informally, we are trying to find a subset Q to query labels that are close to the raw candidates X_U. Although this problem is NP-hard [4], we could obtain an approximate solution efficiently using the greedy iterative approarch:

1) Get $x = \arg\max_{x \in X_T \setminus Q} \min_{x'} p_u(x) \|x - x'\|_2$.
2) Add sample x to the subset Q: $Q = Q \cup \{x\}$.

4.2 Model Adaptation Module $\mathcal{M}_A$

Given the newly queried samples $Q = \{X_Q, y_Q\}$, we have environments $\mathcal{E} = [(X_S, y_S), (X_Q, y_Q)]$. Then we could update the invariant feature Φ via standard environment-based invariant learning. Following [14, 35, 15, 19], we mine the invariance on raw feature level, a simple but general setting. We further obtain Φ through feature selection: $\Phi(X) = M \odot X$.

The objective function of $\mathcal{M}_A$ with $M \in \{0, 1\}^d$ is:

$$\min_{M,f} \sum_{D^e, e \in \mathcal{E}} R(f \circ M | D^e), \quad \text{subject to} \quad f \in \arg\min_{\tilde{f}} R(\tilde{f} \circ M | D^e), \forall e \in \mathcal{E}.$$

However, as the optimization of hard feature selection with binary mask M suffers from high variance, we use the soft feature selection with gates taking continuous value in $[0, 1]$.

LOG is applicable broadly to environment-based invariant learning objectives through the different choices of $\mathcal{M}_A$. In this paper, we choose IGA [13] which has guaranteed to achieve the maximal invariant predictor with respect to given environments.

4.3 Interactive Promotion

Theorem 4.3 (Interactive Promotion)**.** *Given the newly queried samples from $\mathcal{E}_U$, invariant learning module $\mathcal{M}_A$ could promote the invariance set $\mathcal{I}_S$ to ideal $\mathcal{I}_T$. Given the updated invariance set $\mathcal{I}' \subset \mathcal{I}_S$, we have better generalization: $\mathcal{E}'_G \supset \mathcal{E}_G$ and reduced the candidates for actively sampling.*

The core of our LOG framework is the mechanism for $\mathcal{M}_Q$ and $\mathcal{M}_A$ to mutually promote each other. Here we theoretically justify the positive feedback. It indicates that our active exploration could help the reduction of the current invariance set $\mathcal{I}_S$. On the other hand, the better invariance set $\mathcal{I}$ could help to expand the generalizable distributions and reduce the querying candidates.

5 Empirical Study

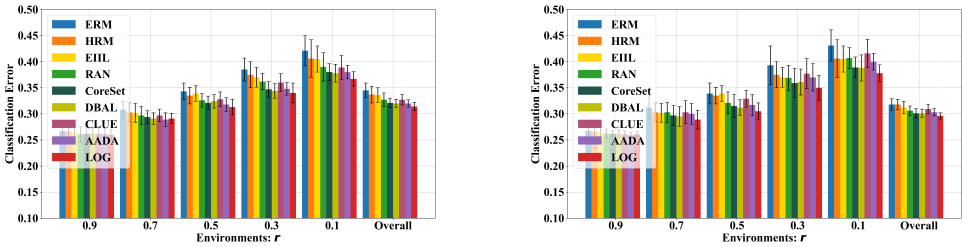
In this section, we provide extensive results to evaluate LOG and compared methods for both benchmark simulation and a series of real-world tasks.

Competing Methods We firstly consider learning only on source: including ERM baseline, and 2 state-of-the-art OOD generalization methods: HRM [19] and EIIL [5], which mine the latent heterogeneity without prior division label $e \in \mathcal{E}_S$. Then we consider active model adaptation, including randomly querying and CoreSet [32] baselines, 3 state-of-the-art active model adaptation methods: AADA [38], CLUE [28], DBAL [6].

We use Logistic Regression and Linear Regression as the base models for classification and regression, respectively. Our feature selection weights M could be derived directly from the parameters of them.

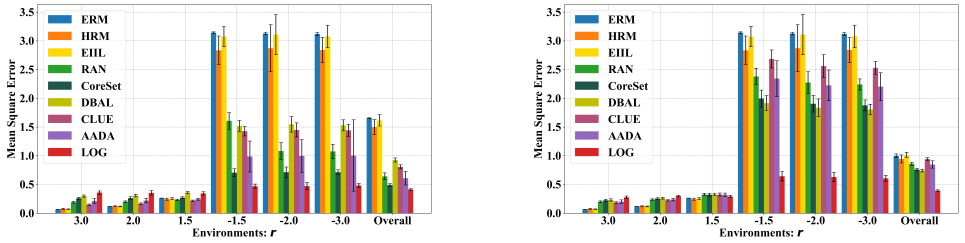
5.1 Simulation Data

Synthetic data are important tools to simulate explainable and controllable distributional shifts. As indicated by [2, 36], it is necessary to introduce such simple but challenging data, which can reflect whether and to what extent an algorithm can resist certain kinds of distributional shifts. Generally speaking, the covariates X are divided into $X = [S, V]^T$, corresponding to the invariant and variant parts inside the data. The $P(Y|S)$ remains invariant across distributions. The $P(Y|V)$ is perturbed with different mechanisms, which brings a distribution shift.



(a) Error under Anti-Causal Effect Shifts

(b) Error under Anti-Causal Effect Shifts (Imbalanced)



(c) Error under Selection Bias Shifts

(d) Error under Selection Bias Shifts (Imbalanced)

Figure 2: Results on varying base distributions (under 10% labeling budgets).

Classification with Anti-Causal Effect Arjovsky et al.(2019) introduce anti-causal relationship to change $P(Y|V)$. Specifically, each distribution is characterized by its bias rate $r \in (0, 1]$, measureing the strength and direction of the spurious correlation between $Y \in \{1, 0\}$ and $A \in \{1, 0\}$. To be detailed, bias rate r represents that for $100 * r\%$ data, $A = Y$, and for the other $100 * (1 - r)\%$ data, $A = 1 - Y$. Then invariant S and variant V are generated as:

$$S|Y \sim \mathcal{N}(Y\mathbf{1}, \sigma_s^2), V|A \sim \mathcal{N}(A\mathbf{1}, \sigma_v^2) \tag{9}$$

We generate 1000 samples from $r_s = 0.9$ as source data D_S, 5000 samples from 5 uniform environments $\mathcal{E}_T$ with $r \in [0.9, 0.7, 0.5, 0.3, 0.1]$ as unlabeled data pool X_T (1000 samples for each r). We carry out the procedure 10 times and report the average results in Figure 2(a).

Regression with Selection Bias Kuang et al.(2020) propose a selection bias mechanism to introduce distributional shifts, and similar settings are also adopted in [19, 18] The data are generated as: $Y = f(S) + \epsilon$ where $\epsilon \perp V$. The selection probability of certain data point (x, y)

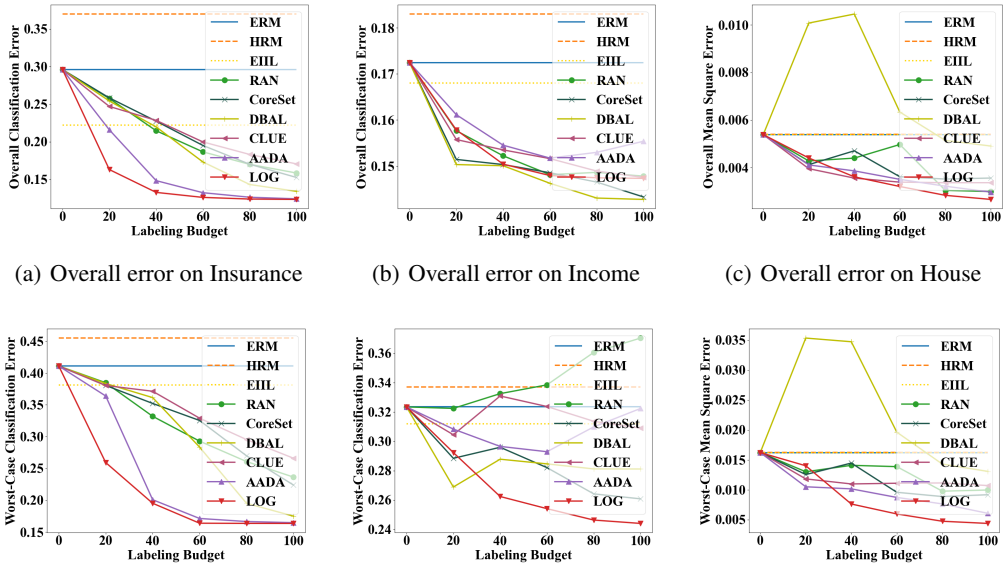
$$P(x, y) = \prod_{v_i \in V} |r|^{-5*|y - \text{sign}(r)*v_i|} \tag{10}$$

where $|r| > 1$. Intuitively, r controls the strengths and direction of the spurious correlation between V and Y. The larger $|r|$ means the stronger spurious correlation between V and Y. $r > 0$ means positive correlation and vice versa. Therefore, we can adopt different r to simulate varying distributions.

We generate 2000 samples from $r_s = 2.0$ as source data D_S, 3000 samples from 6 uniform environments with $r \in \mathcal{E}_T = [3.0, 2.0, 1.5, -1.5, -2.0, -3.0]$ as unlabeled data pool X_T (500 samples for each r). We carry out the procedure 10 times and report the average results in Figure 2(c).

Imbalanced Mixture In the real world, there is a natural phenomenon that empirical data follow a power-law distribution. Here we further simulate an imbalanced situation where the source distribution dominates the target distribution collection $\mathcal{E}_T$. Specifically, we generate half of the target samples from r_S and generate the other from different r. In this case, it is more difficult to query the target-specific samples than in the uniform case. We report the results in Figure 2(b) and 2(d).

From the results, HRM and EIIL could improve worst-case robustness over the ERM baseline. Without sufficient multi-source labeled data, their robustness is still limited because source heterogeneity



(a) Overall error on Insurance (b) Overall error on Income (c) Overall error on House

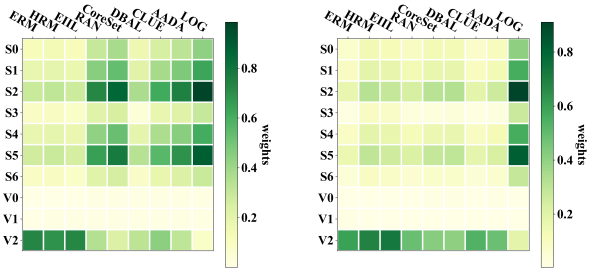
(d) Robustness error on Insurance (e) Robustness error on Income (f) Robustness error on House

Figure 4: Results on real-world tasks. The average results for 10 times produces are reported.

is not sufficient to support varied unseen distributions. In contrast, active adaptation methods which introduce the few labeled data have achieved more significant improvement.

Under the anti-causal shift, all of these active adaptation methods have improved the overall performance of target distributions $\mathcal{E}_T$. Nevertheless, they show significant performance degradation in the base distribution compared to the source distribution. Our LOG has an error which is close to the source error, showing superiority in robustness.

Under selection bias shift, we note that although these adaptation methods have achieved improvement on overall performance, their performance on base distributions (3.0, 2.0, 1.5) close to the source is weaker than source-only methods. To obtain a more clear explanation, we further visualize the dependence of them on invariant variables (S0-S6) and variant variables (V0-V2). One plausible reason is that the source-only methods directly fit the source-specific correlation, overfitting the source distribution. As illustrated in Figure 3, previous methods over-focus on source-specific correlation V2 leads to performance degradation in shifted distributions. In contrast, our method pays more attention to invariant variables (S0-S6) and excludes the variant variables (V0-V2).



(a) Uniform Case (b) Imbalanced Case

Figure 3: Feature importance for each method.

5.2 Real-world Data Sets

We further evaluate our method on three real-world tasks, including car insurance prediction, people income prediction, and house price prediction, with diverse shift types: region, person, and time.

To evaluate both the overall performance and worst-case robustness, we employ the overall error: $\text{error}(f; D_T)$ and worst-case error: $\max_{e \in \mathcal{E}_T} \text{error}(f; D^e)$.

Car Insurance Prediction In this task, we use a real-world dataset for car insurance prediction (Kaggle). It is a classification task to predict whether 381109 persons will buy car insurance based on related information, such as vehicle damage, and vehicle age[3]. We split the dataset into 7 sub-distributions, according to the *region* of these persons. 50% samples from the first region are split in the source labeled data D_S, and the rest is regarded as unlabeled data pool. We report the results under varying labeling budgets in Figure 4(a) and 4(d).

People Income Prediction In this task, we use the Adult dataset to predict personal income levels as above or below \$50000 per year based on personal details. There are 48842 instances in this dataset. we split them into 10 groups according to demographic attributes *sex* and *race*. 50% samples from the first group are split in the source labeled data D_S, and the rest is regarded as unlabeled data pool. We report the results under varying labeling budgets in Figure 4(b) and 4(e).

House Price Prediction In this task, we use a real-world regression dataset of house sales prices from King County, USA[4]. The target variable is the transaction price of the house. Each sample contains 17 predictive variables, such as the built year, number of bedrooms, square footage of the home, etc. Since it is fairly reasonable to assume the relationships between predictive variables and the target vary along the time (for example, the pricing model may change along the time), there exist distributional shifts in the price prediction task concerning the build year of houses. Here, 50% samples of houses built in [1900, 1990] are split in the source labeled data D_S, and the rest samples are regarded as unlabeled data pools. We evaluate model on base distributions through time intervals: [1991, 1995], [1996, 2000], [2001, 2005] and [2006, 2020], to obtain the observations under time shift. We report the results under varying labeling budgets in Figure 4(c) and 4(f).

Analysis From the results, we have the following observations and analyses: The unstable performance of HRM and EIIL indicates the difficulty of generalizing to varied target distributions only by the heterogeneous nature of the source. Randomly querying could perform better than source-only methods at a small labeling cost. In the income task, the source person group dominates the whole distribution, being an imbalanced case. Randomly querying more samples increases the risk of robustness and shows poor worst-case errors. Nevertheless, our LOG has shown stable robustness. CoreSet and CLUE achieve performance gain under distribution shift but are still not satisfying. An interesting phenomenon is that in the early stage of adaptation, DBAL achieved a high improvement in the income task but a significant drop in the house task. One plausible reason is that these samples queried by discrepancy measures are much different from the source data. While they bring rich heterogeneous information, they also increase the risk of the model adaptation part. In contrast, our LOG consistently achieves improvement in overall performance and worst-case robustness, showing the effectiveness of our framework.

5.3 Ablation Study

To evaluate the mutual promotion between two sub-modules $\mathcal{M}_A$ and $\mathcal{M}_Q$, we further make the ablation study here. Specifically, we remove $\mathcal{M}_A$ and $\mathcal{M}_Q$ respectively for comparative experiments. In Table 1, we report the overall error and worst-case error on these real-world tasks. It demonstrates that we could tend to both overall performance and robustness through our interaction between active exploration and invariance exploitation.

Table 1: Overall and worst-case error under 100 labels.

	$\mathcal{M}_A$	$\mathcal{M}_Q$	Insurance	Income	House
O.	√		.131±.008	.154±.003	.296±.029
		√	.128±.005	.153±.014	.344±.205
	√	√	**.123±.000**	**.150±.002**	**.266±.012**
W.		√	.185±.028	.277±.027	.949±.190
		√	.168±.005	.368±.139	.578±.391
	√	√	**.164±.000**	**.235±.006**	**.439±.052**

[3]https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction

[4]https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

6 Conclusion

In this paper, we study the active model adaptation to rectify a known model adapting to a variety of distributions. To our best knowledge, we first introduce the invariant risk minimization principle to guide active adaptation, which leads us to optimal performance and worst-case robustness. Based on the structural causal model assumption, we find the generalization could be significantly improved by expanding the heterogeneity of training data. It motivated us to actively expand the data heterogeneity. We further propose an algorithm LOG that integrates query strategy and invariant model adaptation, with an unbiased estimator to detect the un-generalizable samples. We theoretically justify the mutual promotion relationship between our two sub-modules, resonating with the joint process. A series of empirical studies validate the effectiveness of our algorithm in terms of performance and robustness.

This work focuses on the active model adaptation and provides a promising perspective on label-efficient OOD generalization. Our framework mainly focuses on the raw-level feature selection and the corresponding empirical studies are conducted on tabular tasks. The current proposal is not applicable to the high-dimensional data modality, such as image data. We will put efforts to integrate the power of representation learning capabilities of neural networks and explore broader applications.

Acknowledgement

This research was supported by the National Science Foundation of China (62176118, 61921006), and the Huawei Cooperation Fund. We are grateful for the anonymous reviewers for their helpful comments.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

[2] Benjamin Aubin, Agnieszka Slowik, Martín Arjovsky, Léon Bottou, and David Lopez-Paz. Linear unit-tests for invariance discovery. In *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*, 2021.

[3] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

[4] William Cook, László Lovász, Paul D Seymour, et al. *Combinatorial optimization: papers from the DIMACS Special Year*, volume 20. American Mathematical Soc., 1995.

[5] Elliot Creager, Jörn-Henrik Jacobsen, and Richard S. Zemel. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2189–2200, 2021.

[6] Antoine de Mathelin, Mathilde Mougeot, and Nicolas Vayatis. Discrepancy-based active learning for domain adaptation. *CoRR*, abs/2103.03757, 2021.

[7] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6447–6458, 2019.

[8] Bo Fu, Zhangjie Cao, Jianmin Wang, and Mingsheng Long. Transferable query selection for active domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7272–7281, 2021.

[9] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189, 2015.

[10] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8256–8266, 2018.

[11] Sheng-Jun Huang, Jia-Wei Zhao, and Zhao-Yang Liu. Cost-effective training of deep cnns with active model adaptation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1580–1588, 2018.

[12] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, pages 1675–1685, 2017.

[13] Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. *CoRR*, abs/2008.01883, 2020.

[14] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1617–1626, 2018.

[15] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 4485–4492, 2020.

[16] Hank Liao. Speaker adaptation of context dependent deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7947–7951, 2013.

[17] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In *Advances in Neural Information Processing Systems 34*, pages 6155–6170, 2021.

[18] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernelized heterogeneous risk minimization. In *Advances in Neural Information Processing Systems*, 2021.

[19] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6804–6814, 2021.

[20] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2208–2217, 2017.

[21] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3071–3085, 2019.

[22] Massimiliano Mancini, Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Inferring latent domains for unsupervised deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):485–498, 2021.

[23] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 10–18, 2013.

[24] A. Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. Domain invariant representation learning with domain density transformations. In *Advances in Neural Information Processing Systems 34*, pages 5264–5275, 2021.

[25] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[26] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5102–5112, 2019.

[27] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[28] 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*, pages 8505–8514, 2021.

[29] Harish G. Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 2052–2060, 2016.

[30] Daniel Rejmanek, Parviez R Hosseini, Jonna AK Mazet, Peter Daszak, and Tracey Goldstein. Evolutionary dynamics and global diversity of influenza a virus. *Journal of Virology*, 89(21):10993–11001, 2015.

[31] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, pages 489–511, 2013.

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[33] Jie-Jing Shao, Zhanzhan Cheng, Yu-Feng Li, and Shiliang Pu. Towards robust model reuse in the presence of latent domains. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 2957–2963, 2021.

[34] Jie-Jing Shao, Yunlu Xu, Zhanzhan Cheng, and Yu-Feng Li. Active model adaptation under unknown shift. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1558–1566, 2022.

[35] Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2185–2193, 2020.

[36] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021.

[37] Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime G. Carbonell, and Kun Zhang. Domain adaptation with invariant representation learning: What transformations to learn? In *Advances in Neural Information Processing Systems 34*, pages 24791–24803, 2021.

[38] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 728–737, 2020.

[39] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 4627–4635, 2021.

[40] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 1921.

[41] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2020.

[42] Yu-Jie Zhang, Peng Zhao, Lanjihong Ma, and Zhi-Hua Zhou. An unbiased risk estimator for learning with augmented classes. In *Advances in Neural Information Processing Systems*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative social impacts of your work? [Yes] We provide it in the appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide the complete proofs in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We upload the code as supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify the experiment details about data sets on the Section 5 of the main body. The discussion about hyper-parameters tuning is also provided in the appendix.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5, we report all the results over 10 runs.

(d) Did you include the total amount of computing and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide it in the appendix. In short, we conduct all experiments based on a GPU: RTX 3090.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Broader Impact

In this work, we study the *active learning*, in a real-life scenario where testing distribution shifts from the training. Such a problem widely exists in applications such as finance, self-driving, and biomedical engineering. Active learning methods, including the proposed LOG framework in this paper, aim to obtain more performance gain with limited human-labeling resources. We believe that proper usage of these techniques will lead us to a better society. For example, better active learning techniques could reduce the overhead of the manual labeling process and save computation and storage resources. Better out-of-distribution generalization could promote system deployment in high-stake domains. With that being said, we are also aware that employing these techniques improperly can cause negative impacts, as misprediction is inevitable in most learning algorithms. In particular, we note that when deploying active learning systems in driving-related domains, misprediction (e.g., failure to identify a pedestrian) could lead to a traffic accident. In such domains, these techniques should be used as an auxiliary system, e.g., when it provides driving advice, the driver can take control at any time to maintain safety. As we mentioned above, while there are some risks with these researches, we believe that with proper usage and monitoring, the negative impact of misprediction could be minimized and related techniques could help people live a better life.

B Proofs

B.1 Proof of Theorem 3.6

Theorem B.1. *For a predictor $\Phi^*(X)$ satisfying Assumption 3.3 in the distributions $\mathcal{E}$, the solution $f^*(X) = P(Y|\Phi^*)$ could achieve both performance maximization $R(f^*; \mathcal{D}_T) \leq \min_f R(f; \mathcal{D}_T)$ and worst-case robustness $\max_{e\in\mathcal{E}} R(f^*; \mathcal{D}^e) \leq \min_f \max_{e\in\mathcal{E}} R(f; \mathcal{D}^e)$.*

Proof. We denote the variant and invariant parts of the data as Ψ and Φ^*, respectively. Since optimal Φ^* satisfied distributions invariance, each base-distribution $\mathcal{D}^e$ could be regarded as $P^e([\Phi^*, \Psi], Y) = P(\Phi^*, Y)P^e(\Psi)$.

Considering $f \in \mathcal{F}$ and $f^* = P(Y|\Phi^*)$, we could have $R(f^*; P(\Phi^*, Y)) \leq R(f; P(\Phi^*, Y))$ holds.

Performance Maximization:
Firstly, we would like to prove that for any $f \in \mathcal{F}$ and target distribution $\mathcal{D}_T = \sum_e \lambda^e \mathcal{D}^e$, the following equation holds:

$$R(f; \mathcal{D}_T) \geq R(f^*; \mathcal{D}_T) \tag{11}$$

$$\begin{aligned}
R(f; \mathcal{D}_T) &= \sum_e \lambda^e R(f; \mathcal{D}^e) \\
&= \sum_e \lambda^e \int_{\phi,\psi,y} \ell(f; \phi, \psi, y) p^e(\phi, \psi, y) d\phi d\psi dy \\
&= \sum_e \lambda^e \int_{\phi,y} \int_{\psi} \ell(f; \phi, y) p(\phi, y) d\phi dy p^e(\psi) d\psi \\
&\geq \sum_e \lambda^e \int_{\phi,y} \int_{\psi} \ell(f^*; \phi, y) p(\phi, y) d\phi dy p^e(\psi) d\psi \\
&= \sum_e \lambda^e \int_{\phi,\psi,y} \ell(f^*; \phi, \psi, y) p^e(\phi, \psi, y) d\phi d\psi dy \\
&= \sum_e \lambda^e R(f^*; \mathcal{D}^e) = R(f^*; \mathcal{D}_T).
\end{aligned} \tag{12}$$

Robustness Preservation:
We could directly conclude the robustness through Theorem 2.1 from [19]. For completeness, we give the full proof here.

We would like to prove that for any $f \in \mathcal{F}$, there is e' satisfying the following equation:

$$R(f; \mathcal{D}^{e'}) \geq \max_e R(f^*; \mathcal{D}^e) \tag{13}$$

Let $\bar{e} = \arg\max_e R(f^*; \mathcal{D}^e)$,

$$\begin{aligned}
R(f; \mathcal{D}^{e'}) &= \int_{\phi,\psi,y} \ell(f; \phi, \psi, y) p^{e'}(\phi, \psi, y) d\phi d\psi dy \\
&= \int_{\phi,y} \int_{\psi} \ell(f; \phi, y) p(\phi, y) d\phi dy p^{e'}(\psi) d\psi \\
&\geq \int_{\phi,y} \int_{\psi} \ell(f^*; \phi, y) p(\phi, y) d\phi dy p^{e'}(\psi) d\psi \\
&= \int_{\phi,y} \ell(f^*; \phi, y) p(\phi, y) d\phi \\
&= \int_{\phi,y} \int_{\psi} \ell(f^*; \phi, y) p(\phi, y) d\phi dy p^{\bar{e}}(\psi) d\psi \\
&= \int_{\phi,\psi,y} \ell(f^*; \phi, \psi, y) p^{\bar{e}}(\phi, \psi, y) d\phi d\psi dy \\
&= R(f^*; \mathcal{D}^{\bar{e}}).
\end{aligned} \tag{14}$$

□

B.2 Proof of Theorem 3.7

Theorem B.2. *For distribution $P^{e'}(X^{e'}, Y^{e'})$, if $\Phi(X^{e'}) = \Phi(X^e)_{e \in \mathcal{E}_S}$, $\mathcal{I}_S$ is equal to the invariance set constrained by $\mathcal{E}_S \cup \{e'\}$. The optimal source model f_S could generalize on the distribution $P^{e'}$. The generalize distributions: $\mathcal{E}_G = \{e' | \mathcal{I}_S = \mathcal{I}_{\mathcal{E}_S \cup \{e'\}}\} = \{e' | \Phi(X^{e'}) = \Phi(X^e), e \in \mathcal{E}_S\}$.*

Particularly in the linear structural causal model, we further assume the semantics Z_s takes values in $\mathbb{R}^c$, intervention Z_v^e takes values in $\mathbb{R}^w$, and $S \in \mathbb{R}^{d \times (c+w)}$. Let $\Phi \in R^{d \times d}$, we have:

$$\mathcal{E}_G = \{e' | \mathbb{E}[X^{e'}] = \mathbb{E}[X^e] - x, e \in \mathcal{E}_S, x \in \ker(\Phi)\} \tag{15}$$

where $\dim(\ker(\Phi)) = \dim(\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})) - 1$.

Proof. Denote the invariance set with respect to $\mathcal{E}_S \cup \{e'\}$ as $\mathcal{I}_{\mathcal{E}_S\cup\{e'\}}$. For all $S \in \mathcal{I}_S$, we have $S \in \mathcal{I}_S = \mathcal{I}_{\mathcal{E}_S\cup\{e'\}}$, because the newly queried distribution e' cannot exclude any variables from the current invariance set $\mathcal{I}_S$. Thus, we have the generalizable distributions:

$$\mathcal{E}_G = \{e'|\mathcal{I}_S = \mathcal{I}_{\mathcal{E}_S\cup\{e'\}}\} = \{e'|\Phi(X^{e'}) = \Phi(X^e), e \in \mathcal{E}_S\}.$$

Following the same data generation process, as the structural causal model stated, we have $\Phi(X^e) = \Phi(X^{e'})$ iff $\Phi(\mathbb{E}[X^e]) = \Phi(\mathbb{E}[X^{e'}])$. Furthermore, let us consider the linear structural causal model, the condition could be tailored as:

$$\begin{aligned}\mathcal{E}_G &= \{e'|\Phi(\mathbb{E}[X^{e'}]) - \Phi(\mathbb{E}[X^e]) = 0, e \in \mathcal{E}_S\} \\ &= \{e'|\Phi(\mathbb{E}[X^{e'}] - \mathbb{E}[X^e]) = 0, e \in \mathcal{E}_S\}\end{aligned} \tag{16}$$

It holds when we have $\mathbb{E}[X^{e'}] = \mathbb{E}[X^e]$ or $\mathbb{E}[X^{e'}] - \mathbb{E}[X^e] \in \ker(\Phi)$. The Equation 15 could be conducted.

It demonstrates that the generalization is dependent on the $\mathcal{E}_S$ and Φ. Moreover, the rank of $\ker(\Phi)$ decides the freedom of $\mathcal{E}_G$. We further consider how the $\mathcal{E}_S$ influence the nullspace $\ker(\Phi)$. Consider a basis $\{b_i\}_{i=1}^r$ of the subspace $\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})$, i.e. all of the $\mathcal{E}[X^e]_{e\in\mathcal{E}_S}$ could be combined by $\sum_i^r \lambda_i b_i$. Then we further consider the freedom of the null space of Φ, i.e., the dimension of $\text{span}(\{(b_i - b_j)\}_{i,j\in[1,r]})$. We could conclude the $\dim(\text{span}(\{(b_i - b_j)\}_{i,j\in[1,r]})) = r - 1$. Specifically, let us consider a set of $\{b_1 - b_2, b_1 - b_3, \ldots, b_1 - b_r\}$ with $r - 1$ elements. First, the elements in this set is linear independent. Second, all element in the $\text{span}(\{(b_i - b_j)\}_{i,j\in[1,r]})$ could be linear combined by this set. Thus, we could conclude the nullspace $\ker(\Phi)$ has $\dim(\ker(\Phi)) = \dim(\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})) - 1$.

□

B.3 Proof of Theorem 3.10

Theorem B.3. *Under the linear structural causal model, when we query a distribution $P^{e'}$ where $\Phi(X^{e'}) \neq \Phi(X^e)_{e\in\mathcal{E}_S}$, and update $\Phi \to \Phi'$, $\text{rank}(\Phi') = \text{rank}(\Phi) - 1$ holds. The Φ will converge to the ideal Φ^* at most w steps, where w is the freedom of the intervention variable $Z_v^e \in \mathbb{R}^w$.*

Proof. First, we query a distribution $P^{e'}$ where $\Phi(X^{e'}) \neq \Phi(X^e)_{e\in\mathcal{E}_S}$. As our proposition stated, we have $\Phi\mathbb{E}[X^{e'}] \neq \Phi\mathbb{E}[X^e]_{e\in\mathcal{E}_S}$. We could find the $\mathbb{E}[X^{e'}] \notin \text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})$, otherwises there is $\Phi(X^{e'}) = \Phi(X^e)_{e\in\mathcal{E}_S}$. Thus, $\dim(\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S\cup\{e'\}})) = \dim(\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}_S})) + 1$. For the updated Φ', we have $\text{rank}(\Phi') = \text{rank}(\Phi) - 1$ holds.

Under the linear case, the causal variable $\mathbb{E}[Z_s]$ is consistent in all $e \in \mathcal{E}$ and the space of intervention variable $\text{span}\{\mathbb{E}[Z_v^e]\}_{e\in\mathcal{E}}$ has a freedom w: $Z_v^e \in \mathbb{R}^w$. Due to $X^e = S[Z_s, Z_v^e]$, we have the $\dim(\text{span}(\{\mathbb{E}[X^e]\}_{e\in\mathcal{E}})) \leq w$. Thus, at most w steps, the ideal Φ could remove the influence from variant intervention and recover the consistent causal Z_s. □

B.4 Additional Supplement for Our Proposed Scheme

Here we give an example in Table 2 to illustrate our proposition. Considering a binary classification between *cats* and *dogs*, where each instance contains 3 features, i.e., animal feature X_1, background feature X_2 and the accompanying host feature X_3. Suppose the target distributions $\mathcal{E}_T = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ and the source distributions $\mathcal{E}_S = \{e_1, e_2\}$, then $\mathcal{I}_T = \{X_1\}$ while $\mathcal{I}_S = \{X_1, X_2\}$. The source distributions only tell us X_3 cannot be included in the invariance set but cannot exclude X_2, because of the limitation of $\mathcal{E}_S$. As Theorem B.2 stated, f_S has learned spurious correlation in the source distributions $\mathcal{E}_S$, not enough to achieve maximal invariant predictor in $\mathcal{E}_T$. By contrast, if we could observe instances from $\mathcal{E}_T \backslash \mathcal{E}_G = \{e_5, e_6\}$, $\mathcal{I}_S$ could be corrected to $\mathcal{I}_T = \{X_1\}$.

Table 2: An example for the proposition.

Y	Cats			Dogs		
$\mathcal{E}$	X_1	X_2	X_3	X_1	X_2	X_3
e_1	Cat	Water	Alice	Dog	Grass	Bella
e_2	Cat	Water	Bob	Dog	Grass	Eileen
e_3	Cat	Water	Carol	Dog	Grass	Chalice
e_4	Cat	Water	Diana	Dog	Grass	Ana
e_5	Cat	Grass	Yatoro	Dog	Water	Topson
e_6	Cat	Tree	Jesse	Dog	Water	Ava
$\mathcal{E}_S$	$\{e_1, e_2\}$					
$\mathcal{E}_T$	$\{e_1, e_2, e_3, e_4, e_5, e_6\}$					

B.5 Proof of Theorem 4.2

Theorem B.4 (Estimation Error Bound)**.** *Assume that 1)*$\inf_{g\in\mathcal{G}} R_n(g) \geq \alpha > 0$*; 2)* $\mathcal{G}$ *is closed under negation, i.e.* $g \in \mathcal{G}$ *if and only if* $-g \in \mathcal{G}$*. Then for any* $\delta > 0$*, with probability at least* $1-\delta$*,*

$$\begin{aligned} R(\hat{g}) - R(g^*) \leq\ & 16L\theta\mathfrak{R}_{N_S,\mathcal{D}_S}(\mathcal{G}) \\ & + 6L\mathfrak{R}_{N_T,\mathcal{D}_T}(\mathcal{G}) + 2C\sqrt{\ln(1/\delta)/2} \\ & + 2C\theta\exp(-2(\alpha/C_\ell)^2/(\theta^2/N_S + 1/N_T)) \end{aligned}$$

where $\mathfrak{R}_{N_S,\mathcal{D}_S}(\mathcal{G})$ *and* $\mathfrak{R}_{N_T,\mathcal{D}_T}(\mathcal{G})$ *are the Rademacher complexotoes of* $\mathcal{G}$ *for the sampling of size* N_S *from* $\mathcal{X}_S$ *and of size* N_T *from* $\mathcal{X}_T$*, respectively.*

This theorem ensures that learning of g os consistent as $N_S, N_T \to \infty$, $R(\hat{g}) \to R(g^*)$ in $\mathcal{O}(\theta/\sqrt{N_S} + 1/\sqrt{N_T})$.

Proof. This theorem could be derived directly by Theorem 4 in [12]. □

B.6 Proof of Positive Feedback

Theorem B.5. *Given the newly queried samples from* $\mathcal{E}_U$*, invariant learning module* $\mathcal{M}_A$ *could promote the invariance set* $\mathcal{I}_S$ *to ideal* $\mathcal{I}_T$*. Given the updated invariance set* $\mathcal{I}' \subset \mathcal{I}_S$*, we have better generalization:* $\mathcal{E}'_G \supset \mathcal{E}_G$ *and reduced the candidates for actively sampling.*

Proof. 1) Active query module $\mathcal{M}_Q$ could promote the invariance learning module $\mathcal{M}_A$:

Note the updated invariance set via queried samples Q as $\mathcal{I}'$, we would like to prove $\mathcal{I}_S \supset \mathcal{I}' \supseteq \mathcal{I}_T$. First, we could obtain $\mathcal{I}_S \supset \mathcal{I}'$ because of $\Phi_S(X_S) \neq \Phi_S(X_Q)$. Specifically, for $\Phi_S \in \mathcal{I}_S$, we have $\Phi_S \notin \mathcal{I}'$. Second, X_Q is sampled from $\mathcal{D}_T = \sum_e \lambda^e \mathcal{D}^e$. We could represent the D_Q via the combination of base distributions: $D_Q \sim \sum_e \mu^e \mathcal{D}^e$. Not the Φ in ideal $\mathcal{I}_T$ is invariant for all base distributions. Thus, for any $\Phi \in \mathcal{I}_T$, we have $\Phi \in \mathcal{I}'$, i.e., $\mathcal{I}' \supseteq \mathcal{I}_T$.

2) Invariance learning module $\mathcal{M}_A$ could promote the active query module $\mathcal{M}_Q$:

First, for any $\Phi \in \mathcal{I}_S$, $\Phi(\mathcal{D}_S) = \Phi(\mathcal{D}_g), \forall g \in \mathcal{E}_G$ holds. For any $g \in \mathcal{E}_G$, $\Phi(\mathcal{D}_S) = \Phi(\mathcal{D}_g), \forall \Phi \in \mathcal{I}'$ holds. We have $\mathcal{E}'_G \supseteq \mathcal{E}_G$. Regard the queried samples $[X_Q, y_Q]$ as a observation from distribution q. We have $q \notin \mathcal{E}_G$ and $q \in \mathcal{E}'_G$. Thus, we have $\mathcal{E}'_G \supset \mathcal{E}_G$.

□

C Experimental Details

C.1 Pipeline

As stated in our main paper, our LOG is applicable broadly to environment-based invariant learning objectives through the different choices of $\mathcal{M}_A$. Here we show the experiments using IRM [1] and IGA [13].

The HRM and EIIL are also applicable for different environment-based invariant learning methods. For a fair comparison, we consistently employ IGA here.

We use Logistic Regression and Linear Regression as the base models for classification and regression, respectively. Our feature selection weights M could be derived directly from the parameters of these two linear models.

All of these methods are implemented via Pytorch. We conduct all experiments based on a GPU: Nvidia RTX 3090.

C.2 Evaluation Metrics

To evaluate both the performance and robustness, we report the following measures to evaluate the model.

- Overall metric: metric$(f; D_T)$.
- Worst-case metric: $\max_e$ metric$(f; D^e)$, for $e \in \mathcal{E}_T$.

For classification and regression tasks, we use accuracy (Acc) and mean square error (MSE) as the performance metric, respectively.

C.3 Hyper-parameters Tuning

The proposed framework consists of 1) distribution separation module g, 2) core-set selection and 3) invariant minimization (mask could be joint learning with invariance). Both 1) and 2) do not have hyper-parameters to tune. For the learning of g, we could derive a globally optimal solution without hyper-parameters. For core-set selection, we solve it by a greedy iterative approach without hyper-parameters. For the invariant minimization part, we follow the previous work [13] to tune the hyper-parameters.

C.4 Additional Results of Simulation

C.4.1 Classification with Spurious Correlation

We conduct experiments with different r_S of source data. In each setting, we carry out the procedure 10 times and report the average results. The results are shown in Table 3.

Table 3: Results in classification simulation of different methods with 10% annotating budget.

r (BIAS RATE)	$r_S = 0.80$		$r_S = 0.85$		$r_S = 0.90$	
METHOD	OVERALL ACC ↑	WORST ACC ↑	OVERALL ACC ↑	WORST ACC ↑	OVERALL ACC ↑	WORST ACC ↑
SOURCE	0.668 ± 0.012	0.603 ± 0.020	0.664 ± 0.013	0.596 ± 0.021	0.655 ± 0.014	0.579 ± 0.029
HRM	0.668 ± 0.013	0.599 ± 0.023	0.663 ± 0.015	0.594 ± 0.035	0.663 ± 0.015	0.594 ± 0.035
EIIL	0.671 ± 0.010	0.606 ± 0.020	0.668 ± 0.010	0.601 ± 0.020	0.664 ± 0.013	0.596 ± 0.021
RAN	0.677 ± 0.013	0.619 ± 0.025	0.676 ± 0.012	0.615 ± 0.024	0.673 ± 0.013	0.610 ± 0.027
CORESET	0.683 ± 0.007	0.626 ± 0.014	0.681 ± 0.008	0.623 ± 0.016	0.679 ± 0.009	0.620 ± 0.016
DBAL	0.684 ± 0.008	0.627 ± 0.012	0.683 ± 0.008	0.626 ± 0.011	0.680 ± 0.008	0.620 ± 0.012
CLUE	0.679 ± 0.010	0.620 ± 0.024	0.676 ± 0.011	0.617 ± 0.023	0.673 ± 0.010	0.611 ± 0.023
AADA	0.684 ± 0.008	0.626 ± 0.014	0.682 ± 0.007	0.624 ± 0.010	0.680 ± 0.007	0.620 ± 0.016
OURS (IRM)	$\mathbf{0.691 \pm 0.007}$	$\mathbf{0.640 \pm 0.010}$	$\mathbf{0.689 \pm 0.007}$	$\mathbf{0.640 \pm 0.012}$	$\mathbf{0.688 \pm 0.009}$	$\mathbf{0.637 \pm 0.011}$
OURS (IGA)	0.688 ± 0.007	0.634 ± 0.010	0.687 ± 0.006	0.634 ± 0.006	0.686 ± 0.008	0.630 ± 0.011

From the results, we could observe that: randomly querying (RAN) still performs better than source-only methods. Besides, DBAL and AADA that focus on the hard or representative samples could achieve performance improvement over RAN under the distribution shift condition. Nevertheless, our LOG consistently outperforms these baselines and shows significant improvement in worst-case accuracy.

C.4.2 Regression with Selection Bias

We conduct experiments with different r_S of source data. In each setting, we carry out the procedure 10 times and report the average results. The results are shown in Table 4.

Table 4: Results in regression simulation of different methods with 10% annotating budget.

r (BIAS)	$r_S = 1.7$		$r_S = 2.0$		$r_S = 2.3$	
METHOD	OVERALL MSE ↓	WORST MSE ↓	OVERALL MSE ↓	WORST MSE ↓	OVERALL MSE ↓	WORST MSE ↓
ERM	1.278 ± 0.074	2.434 ± 0.191	1.656 ± 0.007	3.142 ± 0.014	1.857 ± 0.125	3.610 ± 0.394
HRM	1.247 ± 0.069	2.428 ± 0.160	1.497 ± 0.132	2.998 ± 0.344	1.763 ± 0.161	3.610 ± 0.360
EIIL	1.277 ± 0.074	2.491 ± 0.174	1.616 ± 0.098	3.249 ± 0.289	1.853 ± 0.122	3.802 ± 0.310
RAN	0.597 ± 0.044	0.975 ± 0.129	0.641 ± 0.060	1.081 ± 0.147	0.619 ± 0.048	1.050 ± 0.141
CORESET	0.478 ± 0.014	0.679 ± 0.060	0.487 ± 0.026	0.714 ± 0.040	0.491 ± 0.032	0.745 ± 0.111
DBAL	0.914 ± 0.030	1.562 ± 0.080	0.926 ± 0.035	1.606 ± 0.113	0.938 ± 0.036	1.673 ± 0.085
CLUE†	0.684 ± 0.037	1.173 ± 0.113	0.807 ± 0.037	1.449 ± 0.123	0.806 ± 0.082	1.460 ± 0.213
AADA†	0.580 ± 0.055	0.929 ± 0.121	0.610 ± 0.118	1.004 ± 0.261	0.632 ± 0.110	1.072 ± 0.285
OURS (IRM)	0.426 ± 0.018	0.506 ± 0.064	0.414 ± 0.016	0.488 ± 0.036	0.415 ± 0.012	0.510 ± 0.041
OURS (IGA)	$\mathbf{0.425 \pm 0.017}$	$\mathbf{0.499 \pm 0.063}$	$\mathbf{0.413 \pm 0.016}$	$\mathbf{0.481 \pm 0.035}$	$\mathbf{0.414 \pm 0.012}$	$\mathbf{0.503 \pm 0.040}$

From the results, we could observe that: source-only methods suffers from the distribution shift and lead to poor performance under different settings. All of these active adaptation methods have improved the performance of target distributions. Our proposed LOG achieves significant performance improvement in the different settings. Particularly, its worst-case performance is close to overall performance and proves its excellent robustness.

C.4.3 Imbalanced Mixture

Here we also report the numerical statistics of the imbalanced case in Table 5, corresponding to the results of Figure 2(b) and 2(d) in the main paper.

Table 5: Results in imbalanced mixture distributions with 10% annotation budgets.

TASK	SPURIOUS CORRELATION ($r_S = 0.9$)		SELECTION BIAS ($r_S = 2.0$)	
METHOD	OVERALL ACC ↑	WORST ACC ↑	OVERALL MSE ↓	WORST MSE ↓
ERM	0.682 ± 0.011	0.564 ± 0.028	0.999 ± 0.032	3.032 ± 0.152
HRM	0.682 ± 0.010	0.563 ± 0.033	0.944 ± 0.070	3.021 ± 0.256
EIIL	0.688 ± 0.011	0.579 ± 0.027	1.014 ± 0.045	3.275 ± 0.183
RAN	0.694 ± 0.009	0.592 ± 0.019	0.860 ± 0.027	2.444 ± 0.123
CORESET	0.699 ± 0.009	0.610 ± 0.020	0.759 ± 0.022	2.046 ± 0.114
DBAL	0.699 ± 0.008	0.608 ± 0.021	0.741 ± 0.023	1.970 ± 0.108
CLUE	0.691 ± 0.009	0.583 ± 0.025	0.941 ± 0.026	2.755 ± 0.104
AADA	0.697 ± 0.007	0.596 ± 0.011	0.848 ± 0.063	2.398 ± 0.268
OURS (IRM)	$\mathbf{0.706 \pm 0.005}$	$\mathbf{0.623 \pm 0.012}$	0.394 ± 0.014	0.687 ± 0.062
OURS (IGA)	0.704 ± 0.006	0.619 ± 0.013	$\mathbf{0.393 \pm 0.014}$	$\mathbf{0.674 \pm 0.061}$

As we stated in the main paper, an imbalanced case poses a greater challenge to the worst-case generalization of active adaptation. Nevertheless, our LOG still shows superiority in both overall performance and worst-case robustness.

C.5 Details of Real-world Task

In this paper, we also evaluate our method on a series of real-world tasks. They have three common distribution shifts respectively: region shift, person group shift, and time shift. The information of them has been summarized in Table 6.

Task	Goal	# Ins.	# Env.	Shift
Insurance	Classification	381109	7	Regions
Income	Classification	48842	10	Persons
House	Regression	1460	5	Time

Table 6: The basic information of the datasets in our experiments.

C.5.1 Car Insurance Task

In this task, there are 391109 instances from 52 regions. We devide them into 7 groups basd on region id, i.e., [0, 7], [8, 15], [16, 31], [32, 39], [40, 47], [48, 51]. Then we extract 50% instances (14478) from the first group [0, 7] to construct the source labeled data. All of the rest 376631 instances are regarded as the unlabeled data pool.

C.5.2 People Income Task

In this task, we use the Adult dataset to predict personal income levels as above or below 50000 per year based on personal details. There are 48842 instances in this dataset. Following [19], we divide them into 10 base distributions, according to demographic attributes *sex* and *race*. Then we extract 50% instances (14367) from the first group to construct the source labeled data. All of the rest is regarded as the unlabeled data pool. As shown in Table 7, we could find there is an extremely imbalanced ratio in these groups.

Distribution ID	0	1	2	3	4	5	6	7	8	9
# Instances	14368	2377	2308	13027	1002	285	155	517	185	251

Table 7: Size of base distributions in the target distributions.

The maximum imbalanced ratio between base distributions has achieved $92:1$. This situation makes it difficult for active model adaptation to focus on the minor base distribution.

C.5.3 House Prediction Task

In this task, we use a real-world regression dataset of house sales prices from King County, USA. We split the 50% (455) houses built between [1900, 1990] are extracted as the source labeled data, and the rest samples (1005) are regarded as unlabeled data pool. We evaluate model on base distributions through time intervals: [1991, 1995], [1996, 2000],[2001, 2005] and [2006, 2020], to obtain the observations under time shift.

C.5.4 Ablation Study

To evaluate the mutual promotion between two sub-modules $\mathcal{M}_Q$ and $\mathcal{M}_A$, we further make the ablation study here. Specifically, we use randomly querying to replace $\mathcal{M}_Q$ and ERM to repalce $\mathcal{M}_A$, constructing the comparision experiments. As shown in Table 1 (in main paper), it clearly demonstrates that we could tend to both overall performance and robustness through our interaction between active exploration and invariance exploitation.

C.6 Combination with Representation Learning

In this paper, our experiments mainly focus on tabular data, which has widely real-world applications. Here, we also evaluate it on a more complicated modality, images. We conduct the experiments on the colored MNIST benchmark.

Following [1], we build a synthetic binary classification task, where each image is colored either red or green in a way that strongly and spuriously correlates with the class label Y. By construction, the label is more strongly correlated with the color then with the digit, so any algorithm purely minimizing training error will tend to exploit the color. Specially, a binary label Y is assigned to each images according to its digits: $Y = 0$ for digits $[0, 4]$ and $Y = 1$ for digits $[5, 9]$. We further flip Y with different Secondly, we assign the color id $C \in \{0, 1\}$ by flipping Y with different probabilities r and therefore form distributions. In words, a distribution with probability $r = 0.9$ means the sample (x_i, y_i) $\mathcal{D}^r$ has a 0.9 probability that $C = 1 - y_i$ (a strong spurious correlation). In contrast, a distribution with probability $r = 0.1$ means the sample (x_i, y_i) $\mathcal{D}^r$ has a 0.9 probability that $C = y_i$. It leads a poor generalization under distribution shift if model only predicts by color information.

We generate 5000 samples from $r_s = 0.9$ as source data D_S, 10000 samples from 5 uniform base-distributions with $r \in \mathcal{E}_T = [0.9, 0.7, 0.5, 0.3, 0.1]$ as unlabeled data pool X_T (2000 samples for each r). We carry out the procedure 10 times and report the average results in Table 8.

Table 8: Results in color MNIST with varying annotation budgets.

ANNOTATION	1%		10%	
METHOD	OVERALL ACC ↑	WORST ACC ↑	OVERALL ACC ↑	WORST ACC ↑
ERM	0.528 ± 0.003	0.191 ± 0.008	0.528 ± 0.003	0.191 ± 0.008
HRM	0.510 ± 0.005	0.139 ± 0.008	0.510 ± 0.005	0.139 ± 0.008
EIIL	0.507 ± 0.004	0.132 ± 0.008	0.507 ± 0.004	0.132 ± 0.008
RAN	0.537 ± 0.003	0.216 ± 0.008	0.606 ± 0.003	0.382 ± 0.007
CORESET	0.536 ± 0.004	0.212 ± 0.008	0.599 ± 0.004	0.351 ± 0.015
DBAL	0.536 ± 0.003	0.211 ± 0.005	0.596 ± 0.003	0.341 ± 0.016
CLUE	0.537 ± 0.004	0.215 ± 0.008	0.596 ± 0.003	0.339 ± 0.009
AADA	0.538 ± 0.003	0.215 ± 0.007	0.608 ± 0.003	0.354 ± 0.011
OURS (IRM)	0.545 ± 0.005	$\mathbf{0.259 \pm 0.012}$	$\mathbf{0.618 \pm 0.006}$	0.454 ± 0.015
OURS (IGA)	$\mathbf{0.547 \pm 0.006}$	0.255 ± 0.018	0.614 ± 0.004	$\mathbf{0.472 \pm 0.009}$

As for the model architecture, we build an MLP with 2 hidden layers $\{256, 256\}$. Specifically, the architecture of the MLP is (1) linear layer (input dim, 256), (2) ReLu layer, (3) linear layer (256, 256), (4) ReLu layer, (5)linear layer (256, 1). We take the outputs of the second linear layer as the representations $\Phi(X)$, being a replacement of the feature mask for complicated data modality.

From the results, we could find that our framework could easily extend to the deep feature learning implementation and show its significant effectiveness. At different annotation budgets, it could improve both overall performance and worst-case generalization.

In this work, we have shown our framework could take the benefit of invariant learning to provide worst-case performance generalization for active adaptation. On the other hand, our theoretical and experimental results conclude that querying samples via interaction with invariant learning can effectively alleviate the need for SOTA OOD methods for large amounts of labeled data. In the future, we would explore the different representation learning schemes, like CNN, LSTM, and Transformer, adapting to different data modalities.

D Code Repository

The code is available at `https://www.lamda.nju.edu.cn/code_LOG.ashx`. We provide the README.md for them, following the NeurIPS code submission guidelines [5].

[5] https://neurips.cc/Conferences/2022/PaperInformation/CodeSubmissionPolicy