Towards Understanding Variants of Invariant Risk Minimization through the Lens of Calibration

Anonymous authors
Paper under double-blind review

Abstract

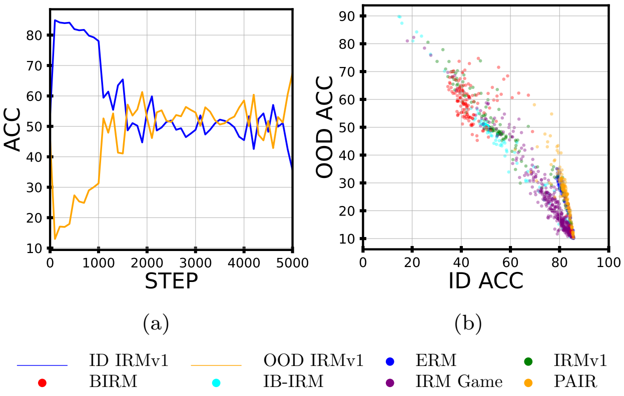
Machine learning models traditionally assume that training and test data are independently and identically distributed. However, in real-world applications, the test distribution often differs from training. This problem, known as out-of-distribution (OOD) generalization, challenges conventional models. Invariant Risk Minimization (IRM) emerges as a solution that aims to identify invariant features across different environments to enhance OOD robustness. However, IRM's complexity, particularly its bi-level optimization, has led to the development of various approximate methods. Our study investigates these approximate IRM techniques, employing the Expected Calibration Error (ECE) as a key metric to measure the extent to which the model has acquired invariant features. ECE, which measures the reliability of model prediction, serves as an indicator of whether models effectively capture environment-invariant features. Through a comparative analysis of datasets with distributional shifts, we observe that Information Bottleneck-based IRM, which condenses representational information, achieves a balance in improving ECE while preserving accuracy relatively. This finding is pivotal, demonstrating a feasible path to maintaining robustness without compromising accuracy. Nonetheless, our experiments also caution against over-regularization, which can diminish accuracy. This underscores the necessity for a systematic approach in evaluating OOD generalization metrics, which goes beyond mere accuracy to address the nuanced interplay between accuracy and calibration. Our code is available at https://anonymous.4open.science/r/IRM_Variants_Calibration-71D0

1 Introduction

In the evolving landscape of machine learning, the importance of out-of-distribution (OOD) generalization (Vapnik, 1991) and calibration (Guo et al., 2017a) cannot be understated, especially in real-world applications and critical scenarios. Traditional machine learning approaches, grounded in the assumption that data are independently and identically distributed (IID), often struggle to cope with OOD scenarios. Additionally, a notable trend in contemporary machine learning models is their tendency towards overconfidence, which undermines the reliability of their confidence estimations. This issue is primarily recognized as a calibration challenge. Responding to these limitations, there has been a surge in research focusing on methodologies like Invariant Risk Minimization (IRM) (Arjovsky et al., 2020). IRM is designed to identify constant features across different environments, facilitating more robust generalization in the face of environmental variations. However, the computational demands of IRM have led to the exploration of more feasible, approximate methods. Despite the development of several IRM variants, their inability to consistently surpass the performance of finely-tuned Empirical Risk Minimization (ERM) models has been observed (Gulrajani & Lopez-Paz, 2020; Zhang et al., 2023a). This performance gap raises critical questions about the inherent deficiencies of these IRM variants. As we showed in Figure 1, even by leveraging IRM variants, there is a trade-off between in-distribution (ID) and OOD accuracy. This implies that models trained by IRM variants failed to obtain invariant features since they sacrificed ID accuracy to achieve better OOD accuracy. From another perspective of IRM, it can also be conceptualized as a specialized form of multi-domain calibration (Wald et al., 2021a). Our study intersects with these findings, examining how IRM variants mimic the original model in calibration. We posit that effective implementation of IRM should lead to reduced loss and a zero expected calibration error (ECE) (Wald et al., 2021a). Through comparative experimentation, we assess various IRM variants in environments marked by distributional shifts, such as diversity and correlation shifts (Ye et al., 2022). Our evaluation focuses not only on accuracy but also on ECE. The results highlight notable disparities in ECE among different methodologies, even when OOD accuracy levels are similar. This suggests that a sole reliance on OOD accuracy for method development might be inadequate. Our observations include a range of calibration behaviors across methods, from overconfidence to underestimation.

Our key contributions are as follows:

- We study discrepancies between IRM formulation and its implemented variants using ECE as a key metric to measure how much the model learned environmental invariant features.
- We reveal that IRM variants with strong regularization typically achieve lower ECE (Figure 3 and Figure 6).
- We observed that Information Bottleneck-based IRM (IB-IRM) lowers ECE across the environments (Figure 5), aligning it more closely with the original objectives of IRM while maintaining OOD accuracy relatively (Figure 2 and Table 2).
- In contrast, for IRMv1, we identify a trade-off between ECE and accuracy, thereby underscoring the compromises necessary for achieving robust generalization and accurate calibration (Figure 4).



(a)

(b)

Figure 1: A graph observed in CMNIST showing the trade-off between ID and OOD accuracy in approximation methods of IRM. Figure (a) visualizes the accuracy of the IRMv1 model ID and OOD, plotted against the training steps on the horizontal axis. The two solid lines are approximately symmetric around an accuracy of 50%. Figure (b) visualizes the ID accuracy (horizontal axis) against the OOD accuracy (vertical axis) for typical approximation methods of IRM. There is a clear trend that as the accuracy improves OOD, the ID accuracy decreases.

Our study, using ECE as a metric, illuminates the gap between the original formulation of IRM and its variants. We recognize a notable trade-off between ECE and accuracy, highlighting the complexities and necessary compromises in striving for robust generalization and precise calibration.

2 Background

2.1 OOD Generalization

In traditional deep neural network scenarios, it is assumed that the learning and test environments are IID. The most commonly used learning algorithm in this assumption has been ERM (Vapnik, 1991), which has been successful in various scenarios. Specifically, ERM aims to minimize the sum of the risk of the model $f : \mathcal{X} \to \mathcal{Y}$ in the set of training environments E_{train}.

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \sum_{e\in E_{train}} R^e(f) \tag{1}$$

where $R^e(f)$ is represented as $\mathbb{E}_{X^e,Y^e}[\ell(f(X^e), Y^e)]$, which is the risk in a environment e.

However, in the real world, the assumption of IID does not always hold, and distribution shifts between training and testing environments can occur (Quinonero-Candela et al., 2008). When the model relies on features that are only effective in the training environment, known as shortcut features, its performance can decrease in the testing environment (Geirhos et al., 2020). This issue is commonly known as OOD generalization and has become an urgent problem recently. Therefore, it is necessary to train robust models that are not dependent on specific environments and can perform consistently across all environments. A model that is robust across the environment is defined as follows.

$$\mathbb{E}[Y^e|f(X^e)] = \mathbb{E}[Y^{e'}|f(X^{e'})], \forall e, e' \in E_{all} \tag{2}$$

where E_{all} represents the set of all possible environments. Equation (2) indicates that, for a robust model, the conditional probability of the label given the model's prediction should be equal between any two environments.

Since it is difficult to access all real-world environments during training, various methods have been proposed to realize invariant models satisfying (eq. (2)) using limited learning environments.

2.2 Calibration

The alignment of a model's predictive confidence with its actual accuracy is extremely important in practical scenarios such as medical image diagnosis and autonomous driving. For example, if there are 100 instances where a model predicts with 80% confidence, then 80 of those predictions should be correct. However, in recent deep neural networks, issues such as increased model capacity have led to problems with miscalibration, and many methods have been proposed to address this calibration issue (Guo et al., 2017a). The correct calibration in all environments is defined for any predictive probability α of f as follows:

$$\mathbb{E}[Y^e|f(X^e) = \alpha] = \alpha, \forall e \in E_{all} \tag{3}$$

Equation (3) implies that the model's conditional predictive probability (in other words, it is called confidence) is always consistent with its actual accuracy.

Furthermore, the quality of uncertainty calibration is often quantified as ECE and is commonly used. It is defined as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |Acc(B_m) - Conf(B_m)| \tag{4}$$

$Acc(B_m)$ and $Conf(B_m)$ are defined as follows

$$\begin{aligned} Acc(B_m) =& \frac{1}{|B_m|} \sum_{b \in B_m} \mathbf{1}\left(\hat{y}_b = y_b\right) \\ Conf(B_m) =& \frac{1}{|B_m|} \sum_{b \in B_m} \hat{p}_b \end{aligned}$$

where the predictive probabilities ranging from 0 to 1 are divided into M parts, $Acc(B_m)$ is the prediction accuracy within the bin B_m and $Conf(B_m)$ is the average of the prediction probability $\hat{p}_b$ (confidence level) for each data in the bin B_m.

2.3 OOD Generalization and Calibration

It has been pointed out that OOD generalization and calibration are interrelated. Notably, eq. (3) can be a sufficient condition for eq. (2), theoretically demonstrating that calibration across all environments leads to OOD generalization. Wald et al. (2021b) capitalize on this observation and propose CLOvE, a method that calibrates models across multiple environments, thereby enabling generalization to unseen data distributions.

Additionally, it has been empirically demonstrated that there is a correlation between a model's accuracy in OOD scenarios and its calibration performance. It has been demonstrated that there exists a correlation between a model's accuracy and its ECE under dataset shift scenarios. Furthermore, techniques such as temperature scaling (Guo et al., 2017b), originally proposed for calibration purposes, have been shown to be effective for OOD generalization (Ovadia et al., 2019; Naganuma & Hataya, 2023).

In this paper, we utilize this fact and hypothesize that models with high calibration performance will likely achieve better OOD generalization. We evaluate the approximation methods of Invariant Risk Minimization (IRM), introduced in Section 4, based on the model's calibration performance to assess their OOD generalization capabilities.

3 Preliminary

3.1 Domain Invariant Representation

One approach to making models more robust in OOD scenarios is through representation learning. Representation learning is a method that teaches the model to learn effective data representations from raw data for better performance in predictions. Traditionally, features of data were designed by human, but representation learning reduces this necessity and can also compress the dimensions of input data, enabling faster computations.

In the context of OOD scenarios, there is an application for representation learning, specifically aimed at extracting environment-invariant features from raw data and reducing environment-specific spurious features to achieve more robust predictions. For example, in the task of classifying animals from images, if the model learns to associate parts of an animal with extraneous elements, it may lead to high performance within the training domain but fail to make robust predictions when the background distribution shifts in the test domain. However, if the model discards spurious features (e.g., background) from the raw data, it can consistently make robust predictions across all environments. One particularly well-known method that applies representation learning to OOD generalization scenarios is Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2016). This approach uses adversarial training to adjust the learning process so that it is impossible to discern from the extracted features which environment the input comes from, thereby aiming to extract invariant features.

3.2 Invariant Learning

In the context of applying representation learning to OOD generalization, the method we focus on in this paper is IRM (Arjovsky et al., 2020). IRM is proposed with the aim of extracting environment-invariant features from input data to enable consistent predictions across any environment. IRM is defined with the purpose of realizing (eq. (2)) as follows.

$$\begin{aligned} &\min_{\substack{\Phi:\mathcal{X}\to\widehat{\mathcal{H}} \\ \omega:\widehat{\mathcal{H}}\to\widehat{\mathcal{Y}}}} \sum_{e\in E_{train}} R^e(\omega\circ\Phi) \\ &\textit{subject to} \quad \omega \in \arg\min_{\bar{\omega}:\widehat{\mathcal{H}}\to\widehat{\mathcal{Y}}} R^e(\bar{\omega}\circ\Phi), \forall e\in E_{train} \end{aligned} \tag{5}$$

Here, $\mathcal{X}$ represents the input space of each data point, $\widehat{\mathcal{H}}$ denotes the extracted invariant feature space, and $\mathcal{Y}$ is the output space of the model. IRM defines the function as $f = \omega \circ \Phi$, where Φ maps the input data to $\widehat{\mathcal{H}}$, and based on these extracted invariant features, the final prediction is made by ω . IRM aims to achieve consistent predictions as a form of OOD generalization.

The formulation in eq. (5) represents a bi-level optimization problem, which is challenging to solve exactly. As a result, various implementable approximation methods have been proposed. Many variants add various constraints to the IRM objective and approximate the bi-level optimization problem as a single-level problem to make it more tractable (Arjovsky et al., 2020; Ahuja et al., 2022; Chen et al., 2022; Ahuja et al., 2020; Lin et al., 2022). Additionally, Zhang et al. (2023b) propose an improved variant of IRM Game, another IRM variant, that retains the bi-level optimization framework during training. However, despite numerous contributions to the development of approximation methods for IRM, a method that generally surpasses well-tuned ERM in terms of OOD generalization accuracy has not yet been proposed. Therefore, it is essential to understand why these approximation methods have not succeeded and what approaches might be more suitable for addressing the approximation problem.

While the original IRM formulation poses implementation challenges, an alternative method has also been proposed to address this issue. IRM aims to learn invariant features by keeping $\mathbb{E}[Y^e|f(X^e)]$ constant across environments. However, Huh & Baidya (2023) introduced a complementary notion of invariance called MRI, which seeks to conserve the label-conditioned feature expectation $\mathbb{E}[f(X^e)|Y^e]$ across environments, in contrast to the original IRM formulation.

4 Related Works

Methods	Featurizer	Predictor	Linearity of ω	Information Bottleneck	Game theory	Ensemble	Bayesian inference
IRM	Φ	ω	✗	✗	✗	✗	✗
IRMv1	$\omega \cdot \Phi$	1.0	✓	✗	✗	✗	✗
IB-IRM	$\omega \cdot \Phi$	1.0	✓	✓	✗	✗	✗
PAIR	$\omega \cdot \Phi$	1.0	✓	✗	✓	✗	✗
IRM Game	Φ	ω_{ens} [1]	✗	✗	✓	✓	✗
BIRM	$\omega \cdot \Phi$	1.0	✓	✗	✗	✗	✓

Table 1: Approximation methods of IRM

4.1 IRMv1

IRMv1 (Arjovsky et al., 2020) constrains ω in eq. (5) as a linear mapping and allows the transformation $\Phi' = (\omega \cdot \Phi)$, $\omega' = 1.0$. The following one-variable optimization problem can be relaxed and computed.

$$\min_{\Phi':\mathcal{X}\rightarrow\mathcal{Y}} \sum_{e\in E_{train}} R^e(\Phi') \ + \ \lambda \parallel \nabla_{\omega'|\omega'=1.0} R^e(\omega' \cdot \Phi') \parallel^2 \tag{6}$$

4.2 Information Bottleneck based IRM(IB-IRM)

While IRMv1 has been successful with certain OOD generalization datasets such as CMNIST, it has been noted that if not enough invariant features are included in each environment in the training data, robust prediction is lost. Therefore, IB-IRM (Ahuja et al., 2022) aims to improve IRMv1 by compressing the entropy of the feature extractor Φ' using the information bottleneck method, thereby preventing excessive reliance on features that adversely affect the invariant prediction. The variance of the features extracted from each data by Φ' is added as a regularization term.

$$\lambda \parallel \nabla_{\omega'|\omega'=1.0} R^e(\omega' \cdot \Phi') \parallel^2 + \gamma Var(\Phi') \tag{7}$$

4.3 Pareto IRM(PAIR)

In general, there is a trade-off between ERM and OOD generalization, and PAIR (Chen et al., 2022) focuses on the need to properly manage that trade-off when generalizing a model out of distribution. To achieve more robust models, PAIR is designed to find their Pareto optimal solutions in terms of multi-objective

[1] $\omega_{ens} = \frac{1}{|E_{train}|}\left(\sum_{e\in E_{train}} w^e\right)$

optimization. In fact, it is implemented as a multi-objective optimal problem for ERM, IRMv1 penalties, and VREx (Krueger et al., 2021) penalties.

$$\min_{\Phi':\mathcal{X}\to\mathcal{Y}} (\mathcal{L}_{ERM}, \mathcal{L}_{IRMv1}, \mathcal{L}_{VREx}) \tag{8}$$

where

$$\mathcal{L}_{ERM} = \sum_{e\in E_{train}} R^e(\Phi')$$

$$\mathcal{L}_{IRMv1} = \| \nabla_{\omega'|\omega'=1.0} R^e(\omega' \cdot \Phi') \|^2$$

$$\mathcal{L}_{VREx} = Var(\{R^e(\Phi')\}_{e\in E_{train}})$$

4.4 IRM Game

IRM Game (Ahuja et al., 2020) aims to achieve Nash equilibrium regarding inference accuracy by introducing a game-theoretic framework to the IRM between each environment. Each environment e in the training data is assigned its own predictor ω^e, each of which is trained to be optimal in each environment, and the final ω_{ens} is the ensemble of all predictors $\frac{1}{|E_{train}|}\sum_{e\in E_{train}} \omega^e$. By learning environment-specific predictors, the limitation on linearity made in IRMv1 is eliminated, and implementation closer to eq. (5) is aimed at.

$$\min_{\substack{\Phi:\mathcal{X}\to\widehat{\mathcal{H}} \\ \omega_{ens}:\widehat{\mathcal{H}}\to\widehat{\mathcal{Y}}}} \sum_{e\in E_{train}} R^e(\omega_{ens} \circ \Phi) \tag{9}$$

$$s.t.\ \omega^e \in \arg \min_{\bar{\omega}^e:\widehat{\mathcal{H}}\to\widehat{\mathcal{Y}}} R^e\left\{\bar{\omega}^e_{ens} \circ \Phi\right\}, \forall e \in E_{train}$$

where $\bar{\omega}^e_{ens} = \frac{1}{|E_{train}|}\left(\bar{\omega}^e + \sum_{e'\neq e} w^{e'}\right)$

4.5 Bayesian IRM(BIRM)

In BIRM (Lin et al., 2022), IRMv1 may overfit the training environment if the training data are insufficient, so we introduced a Bayesian estimation approach there. Noting that the posterior probability $p(\omega^e|\Phi(X^e), Y^e)$ is invariant in all environments given the data mapped by Φ and the correct label if the model is able to acquire invariant features, the penalty of IRMv1 is modified. By using Bayesian estimation instead of point estimation, they aim to improve the generalization performance of the model by taking into account the uncertainty of the data.

Table 1 presents an overview of IRM and its approximation methods.

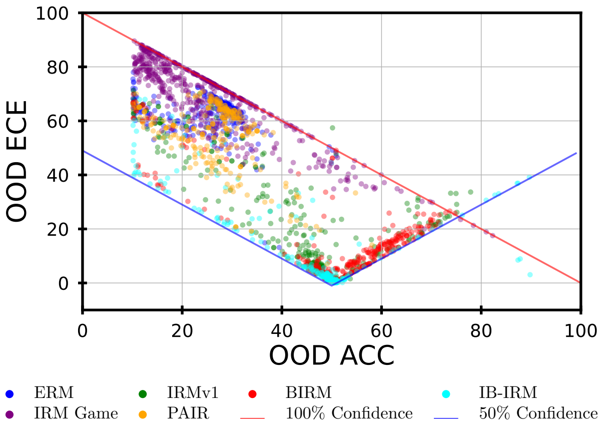


Figure 2: Comparison of OOD accuracy (horizontal axis) and OOD ECE (vertical axis) in CMNIST. The red solid line represents the theoretical values when all the model's predictive probabilities are 100%, and the blue solid line represents the case when all predictive probabilities are 50%. IRMv1, IB-IRM, and BIRM, which achieve relatively high inference accuracy, are distributed near the blue line. This indicates that these models mitigate reliance on non-invariant features by making predictions more ambiguous.

5 Experiments

Although the approximation methods in the previous chapter have achieved some success in terms of OOD accuracy on OOD datasets (Arjovsky et al., 2020; Ahuja et al., 2022; Chen et al., 2022; Ahuja et al., 2020; Lin et al., 2022), it is not known how invariant features they actually acquire and how robust they are to unknown environments. Therefore, we evaluated them by comparing their calibration performance with ECE.

5.1 Setting

We compared ECE of each model on four different OOD-datasets. Specifically, the datasets used were Colored MNIST (CMNIST) (Arjovsky et al., 2020), Rotated MNIST (RMNIST) (Ghifary et al., 2015), PACS (Li et al., 2017), and VLCS (Fang et al., 2013), sourced from the DomainBed benchmark (Gulrajani & Lopez-Paz, 2020). These datasets can be distinguished based on the type of distribution shift they represent (Ye et al., 2022). The distribution shift in CMNIST is known as the correlation shift, where the conditional probability of labels given spurious features changes across environments. The other three datasets—RMNIST, PACS, and VLCS—are classified under diversity shift, where the prior probability of spurious features changes across environments.

For optimization, Adam (Kingma & Ba, 2015) was used consistently across all models, and the tuning of learning rates and hyperparameters for each method was conducted in accordance with their respective papers.

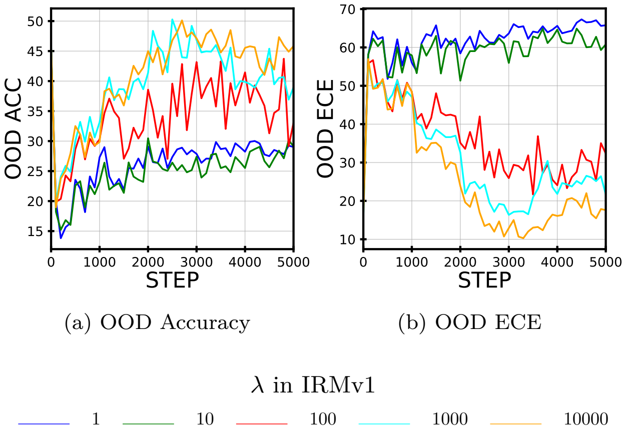
5.2 Correlation Shift

5.2.1 Dataset

CMNIST is a dataset based on MNIST, containing 70,000 dimension examples (1, 28, 28). It is adapted for a binary classification task where digits less than 5 are labeled as class 0, and those 5 or above as class 1. Spurious correlations are added by assigning colors (red or green) to the digits, with the proportion of the two colors varying in each class depending on the environment. Denoting the proportion of green in class 0 as environment e, then the training environments are set as $e = [10\%, 20\%]$, and the test environment as $e = [90\%]$.

5.2.2 Results

Figure 2 presents a scatter plot showing the relationship between OOD accuracy and ECE for multiple IRM variants, each trained with multiple hyperparameters. The red solid line represents the theoretical value when all model's predictive probabilities are 100%, indicating very high confidence. In contrast, the blue solid line represents the case when all predictive probabilities are 50%, indicating very low and ambiguous confidence. The methods with relatively high accuracy (IRMv1(green), IB-IRM(light blue), BIRM(red)) align with the blue solid line, showing that they can overcome excessive fitting to non-invariant features by making predictions more ambiguous. On the other hand, ERM(blue) and IRM Game(purple) tend to be distributed near the red solid line, indicating predictions with relatively high confidence. PAIR(yellow) is distributed between the two solid lines, demonstrating intermediate characteristics between the two aforementioned groups.



(a) OOD Accuracy

(b) OOD ECE

Figure 3: Comparison of the model's OOD ECE in CMNIST, with various values of λ in the formulation of IRMv1 eq. (6). It was observed that increasing λ, thereby increasing the regularization penalty in IRMv1, results in a lower ECE for the model in the OOD environment.

Figure 3 visualizes how the penalty introduced in IRMv1 for the purpose of OOD generalization actually affects the ECE. It specifically compares different values of λ in IRMv1's formulation eq. (6), demonstrating that as the impact of IRMv1's penalty increases, the ECE decreases, suggesting better calibration.

5.3 Diversity Shift

5.3.1 Datasets

The details for RMNIST, PACS, and VLCS are as follows:

RMNIST: RMNIST is a dataset based on MNIST, containing 70,000 examples of the same dimension (1, 28, 28) and includes 10 classes. The environments are distinguished by the rotation angle of the digits, denoted as e. The training environments are set at $e = [15°, 30°, 45°, 60°, 75°]$, while the test environment is at $e = [0°]$.

PACS: PACS is a dataset containing 9,991 examples with dimensions (3, 224, 224) and includes 7 classes. The environments are distinguished by four different styles of images. Specifically, the training environments are set as $e = [Cartoons, Photos, Sketches]$, and the test environment is set as $e = [Art]$.

VLCS: VLCS is a dataset containing 10,729 examples with dimensions (3, 224, 224) and includes 5 classes. There are four types of environments. The training environments are set as $e = [Caltech101, LabelMe, SUN09]$, and the test environment is set as $e = [VOC2007]$.

5.3.2 Results

Considering that a model's OOD accuracy affects its calibration (Ovadia et al., 2019), a comparative evaluation of ECE combined with OOD accuracy was conducted for a fairer comparison of ECE. A threshold was set for validation accuracy in the training environment, and upon reaching this threshold, early stopping of training was implemented to create conditions of comparable OOD accuracy.

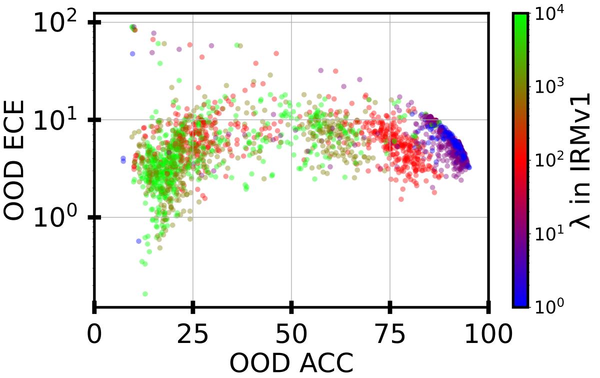
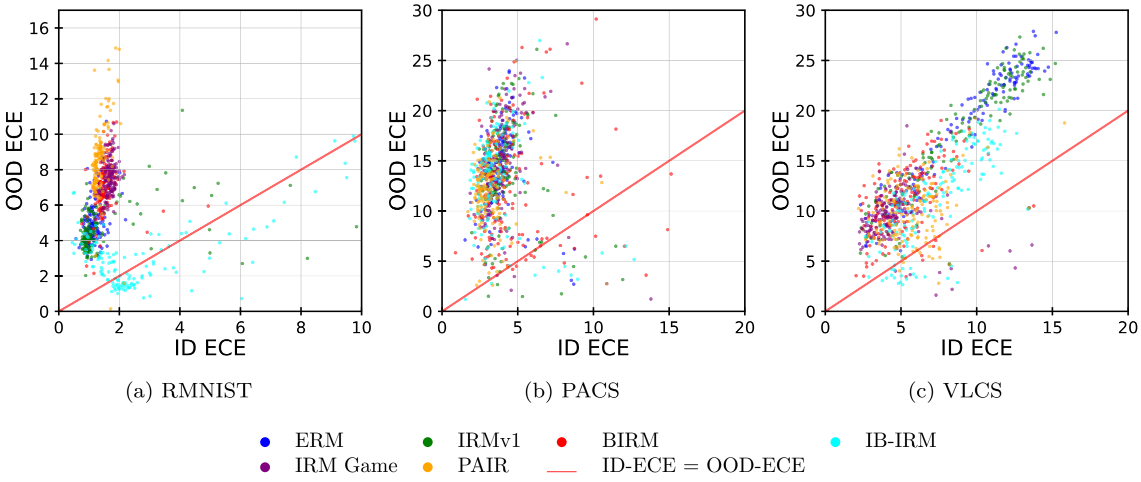


Figure 4: The relationship between OOD accuracy and ECE varies with the λ parameter of IRMv1 in the RMNIST dataset. As the color shifts from blue to green, λ increases, indicating stronger regularization in the model. Points closer to blue, situated on the middle right of the figure, correspond to intermediate ECE and high accuracy. In contrast, points nearer to green, located towards the lower left, indicate better ECE but lower accuracy. This suggests that in more practical datasets, such as those with diversity shifts, stronger regularization may improve ECE, but this often comes at the cost of a significant decrease in accuracy.

The average results from three different random seeds under the aforementioned conditions are presented in Table 2, where although all methods achieved a similar average OOD accuracy, notable differences were observed in OOD ECE. Specifically, the ECE of IB-IRM was, on average, lower than that of any other method. Additionally, Figure 5 visualizes the relationship between ECE in the training and test environments, with the red solid line representing cases where ECE is equal in both environments. Ideally, models should be distributed near this red line from the perspective of cross-environmental calibration of uncertainty. Specifically, in Figure 5a and Figure 5c, IB-IRM is distributed near the red solid line compared to other methods, suggesting that it does not overfit to the training environment and exhibits more ideal feature learning.

It was also suggested that regularization in models such as IRMv1 and IBIRM, while reducing OOD ECE in cases of diversity shift, can adversely affect accuracy. Figure 4 illustrates the relationship between OOD accuracy and ECE in RMNIST when varying the magnitude of λ in IRMv1. As regularization strengthens, the model's ECE improves; however, this is accompanied by a drastic deterioration in accuracy, indicating a trade-off between the two.



(a) RMNIST

(b) PACS

(c) VLCS

Figure 5: Comparison of the relationship between ECE in the training environment (horizontal axis) and the test environment (vertical axis). The red solid line represents the case where the ECE is equal in both environments. It was observed that IB-IRM (in light blue) is distributed near the red solid line, indicating a tendency not to overfit to the training environment compared to other methods.

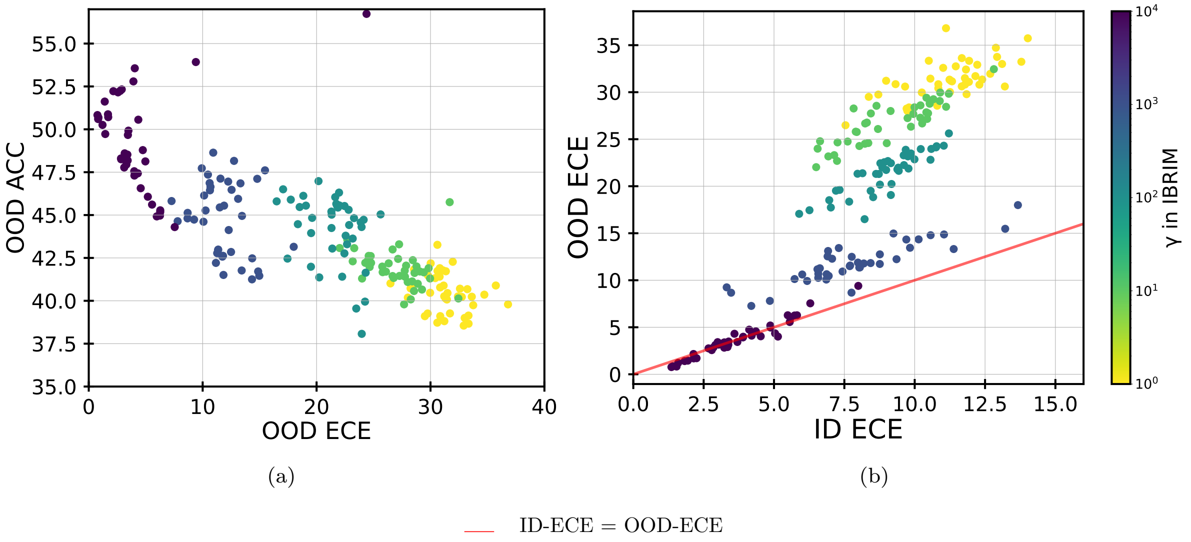
		ERM	IRMv1	IB-IRM	BIRM	IRM Game	PAIR
OOD ACC	RMNIST	91.1±1.0	91.9±1.0	90.9±0.9	89.1±0.4	90.8±1.0	87.4±0.8
	PACS	73.6±2.0	74.9±1.9	75.9±0.9	71.5±1.7	73.2±1.3	76.3±1.2
	VLCS	70.3±1.0	70.5±1.2	69.4±0.7	68.8±0.8	69.9±1.2	71.1±1.2
	Avg.	78.3±1.3	79.1±1.4	78.8±1.12	76.5±0.8	78.1±1.2	78.3±1.1
OOD ECE	RMNIST	5.27±1.11	3.98±0.69	**1.58±0.44**	6.93±0.56	7.17±0.98	7.88±0.79
	PACS	**11.99±1.86**	13.35±1.45	12.85±0.60	13.73±1.93	12.50±1.45	12.46±0.95
	VLCS	12.53±1.42	11.90±1.26	**7.84±1.55**	10.57±1.81	8.70±0.93	10.34±1.11
	Avg.	9.93±1.46	9.74±1.13	**7.42±0.86**	10.41±1.43	9.46±1.12	10.23±0.95

Table 2: Comparison of OOD accuracy and ECE across RMNIST, PACS, and VLCS datasets. Following hyperparameter tuning for each method, measurements were taken across three different random seeds, presenting the mean and variance for each metric. It was observed that on average, IB-IRM exhibits a lower ECE while maintaining comparable inference performance.

6 Information Bottleneck in Calibration

In our experiments, we observe that IRM with information bottleneck is effective in calibration, as shown in Figure 2 and Figure 5. Therefore, in this section, we provide the results of the ablation study for IBIRM.

First, we investigated the impact of the penalty term related to the information bottleneck in IBIRM on calibration. Figure 6 visualizes various evaluation metrics in CMNIST when altering the strength of the information bottleneck in IBIRM, that is, the magnitude of γ in eq. (7). As shown in panel (a), which visualizes this based on OOD ECE and OOD Accuracy, we confirm the clear trend that strengthening the regularization of γ improves both accuracy and ECE. Panel (b) shows the relationship between ECE in ID and OOD. Increasing γ aligns both ECEs towards the red line, representing equality for conditional probability. In other words, this constitutes a successful calibration across environments. These findings suggest that not only is the IRMv1's regularization term aimed at OOD scenarios, but the information bottleneck also further promotes the learning of invariant features.



(a)

(b)

Figure 6: The figures illustrate the impact of the information bottleneck on calibration with IBIRM on CMNIST task. Panel (a) displays the relationship between ECE and Accuracy in the OOD context, investigated by varying the coefficient γ of the information bottleneck penalty in the IBIRM formulation (eq. (7)). As the value of γ increases, both metrics show improvement. Panel (b) visualizes the values of ECE in both ID and OOD. Same as (a), altering the value of γ shows that the larger the value, the more the data aligns with the red line, which indicates equality between the two ECEs. With a sufficiently large γ, the points are almost perfectly distributed along the red line, indicating successful calibration across multiple environments.

Additionally, we demonstrate the validity of our results based on the theoretical principles of the information bottleneck. In OOD scenarios, the objective of the information bottleneck is to minimize the mutual information $I(X; \Phi(X))$ between the input X and the intermediate representation $\Phi(X)$ while preserving the mutual information $I(\Phi(X); Y)$ between $\Phi(X)$ and the output Y. Intuitively, this means learning $\Phi(X)$ to reduce as much information from X as possible while retaining as much information about Y. When $\Phi(X)$ is a deterministic transformation of X, the entropy $H(\Phi(X))$ can be used to minimize $I(X; \Phi(X))$ (Kirsch et al., 2021). In the context of IRM, which aims to learn invariant features, keeping $H(\Phi(X))$ small encourages Φ to capture invariant features X_{inv} from X and discard spurious features X_{sup}. It should be noted that Ahuja et al. (2022) theoretically demonstrated, using realistic SEMs, that minimizing $H(\Phi(X))$ in IRM leads to Φ discarding spurious features and grasping invariant features. Our results empirically support the effectiveness of applying the information bottleneck method to invariant feature learning.

7 Discussion and Conclusion

In this paper, we focus on IRM, which is challenging to implement due to the difficulties of bi-level optimization. We conducted comparative evaluations using ECE, a computationally feasible metric for calibration, to understand its approximation methods better and measure how much the model learned environmental invariant features.

Our experimental results suggested that IB-IRM, which applies the information bottleneck method to IRM, is superior in calibration when we compared models with almost the same accuracy (Figure 2 and Table 2).

Another finding from our empirical results is that regularization penalties introduced for OOD generalization effectively achieve lower ECE. However, in practical datasets (under diversity shift), we observe a trade-off between OOD ECE and accuracy. Although IB-IRM mitigates this tradeoff compared to IRMv1 and other variants; carefully balancing these two metrics is important in aiming for OOD generalization.

As a future challenge, it has become clear that the degree of invariant feature learning cannot be fully measured by accuracy on OOD data alone, necessitating the establishment of systematic generalization metrics and measurement methods. Furthermore, a better understanding of the relationship between OOD generalization and uncertainty calibration, as well as the development of novel approximation methods for IRM based on this understanding, is anticipated. In particular, gaining a deeper understanding of the impact of the information bottleneck on OOD generalization and calibration will be key to addressing the issues of OOD generalization and calibration. A more comprehensive grasp of how the information bottleneck principle affects the model's ability to generalize to unseen environments and maintain well-calibrated predictive uncertainties will be crucial in developing more robust and reliable models.

References

K Ahuja, K Shanmugam, KR Varshney, and A Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

K Ahuja, E Caballero, D Zhang, JC Gagnon-Audet, Y Bengio, I Mitliagkas, and I Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *arXiv preprint arXiv:2106.06607*, 2022.

Martin Arjovsky, L'eon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2020.

Y Chen, K Zhou, Y Bian, B Xie, B Wu, Y Zhang, K Ma, H Yang, P Zhao, B Han, et al. Pareto invariant risk minimization: Towards mitigating the optimization dilemma in out-of-distribution generalization. *arXiv preprint arXiv:2206.07766*, 2022.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

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.

Robert Geirhos, J"orn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017a.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017b. URL https://proceedings.mlr.press/v70/guo17a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Dongsung Huh and Avinash Baidya. The missing invariance principle found – the reciprocal twin of invariant risk minimization, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Andreas Kirsch, Clare Lyle, and Yarin Gal. Unpacking information bottlenecks: Unifying information-theoretic objectives in deep learning, 2021.

D Krueger, E Caballero, J-H Jacobsen, A Zhang, J Binas, D Zhang, R Le Priol, and A Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2021.

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.

Y Lin, H Dong, H Wang, and T Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16021–16030, 2022.

Hiroki Naganuma and Ryuichiro Hataya. An empirical investigation of pre-trained model selection for out-of-distribution generalization and calibration. *arXiv preprint arXiv:2307.08187*, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in neural information processing systems*, volume 32, 2019.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In *Advances in Neural Information Processing Systems*, 2021a.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=XWYJ25-yTRS.

Nanyang Ye et al. 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*, 2022.

Yihua Zhang, Pranay Sharma, Parikshit Ram, Mingyi Hong, Kush Varshney, and Sijia Liu. What is missing in irm training and evaluation? challenges and solutions. *arXiv preprint arXiv:2303.02343*, 2023a.

Yihua Zhang, Pranay Sharma, Parikshit Ram, Mingyi Hong, Kush Varshney, and Sijia Liu. What is missing in irm training and evaluation? challenges and solutions, 2023b.

Appendix

A Details of experiments

A.1 Implementation of IRM Variants

This section provides a detailed description of each method used in the experiments.

A.1.1 IRMv1

The learning process was conducted using the loss calculated based on the formulation shown in 6. The hyperparameter λ is used to adjust the impact of the IRMv1 penalty.

A.1.2 IBIRM

The learning process was performed using the loss calculated based on the formulation shown in 7. The hyperparameters λ and γ are used to adjust the influence of the IRMv1 penalty and the information bottleneck penalty, respectively.

A.1.3 IRM Game

IRM Game has two variants: F-IRM Game and V-IRM Game. The difference between them lies in whether the featurizer Φ is fixed or not. F-IRM Game fixes $\Phi = I$ as an identity matrix, while V-IRM Game considers a variable Φ. For each dataset, the variant that achieved higher performance in terms of accuracy was selected.

A.1.4 PAIR

PAIR employs a weighted average of ERM loss, IRMv1 loss, and VREx loss as the final loss for learning. When calculating the weights, a hyperparameter called $preference$ is used to adjust the scaling of each loss.

A.1.5 BIRM

In BIRM, when calculating the IRMv1 loss, Monte Carlo sampling is used to sample N times from the posterior distribution of the parameters, and the average of these samples is used to compute the loss. This approach takes into account the uncertainty of the parameters when calculating the loss. The standard deviation of the Gaussian noise added to the parameters during sampling is treated as a hyperparameter denoted as $birm\_sd$. In all experiments, N was set to 5.

A.2 Datasets

A.2.1 Domainbed

We conducted experiments using Colored MNIST (CMNIST) (Arjovsky et al., 2020), Rotated MNIST (RMNIST) (Ghifary et al., 2015), PACS (Li et al., 2017), and VLCS (Fang et al., 2013) datasets from DomainBed (Gulrajani & Lopez-Paz, 2020). We split each dataset into training and validation sets, with 80% used for training and the remaining 20% for validation. The environment partitions were as follows:

- CMNIST: $E_{train} = [10\%, 20\%]$, $E_{test} = [90\%]$
- RMNIST: $E_{train} = [15°, 30°, 45°, 60°, 75°]$, $E_{test} = [0°]$
- PACS: $E_{train} = [Photo, Painting, Sketch]$, $E_{test} = [Art]$
- VLCS: $E_{train} = [Caltech101, LabelMe, SUN09]$, $E_{test} = [VOC2007]$

A.2.2 Hyperparameters

In the experiments, we set the batch size to 256 for CMNIST, 128 for RMNIST, and 16 for PACS and VLCS. Grid search was performed on the learning rate for all experiments, with values of [1e-4, 5e-4, 1e-3, 5e-3]. For the hyperparameters specific to each approximation method, grid search was conducted as shown in Table 3.

Parameter	
λ	[1, 1e1, 1e2, 1e3, 1e4]
γ	[1, 1e1, 1e2, 1e3, 1e4]
$preference$	[0, 1, 2, 3, 4]
$birm\_sd$	[5e-2, 1e-1, 2e-1]

Table 3: Hyperparameters for each method

Model selection was performed using an oracle based on accuracy in the test environment for Correlation shift, while for Diversity shift, it was based on validation accuracy in the training environment. Correlation shift is known to be a relatively challenging task(Ye et al., 2022) as it involves a shift in the conditional probability $P[Y|X]$, where X represents the input data and Y represents the corresponding ground truth labels. As shown in Figure 1(b), there exists a trade-off between accuracy in the training and test environments for all methods, necessitating oracle model selection. In contrast, for Diversity shift, there is a correlation between ID and OOD accuracy across all methods as shown in Figure 9, allowing for model selection based on performance in the training environment.

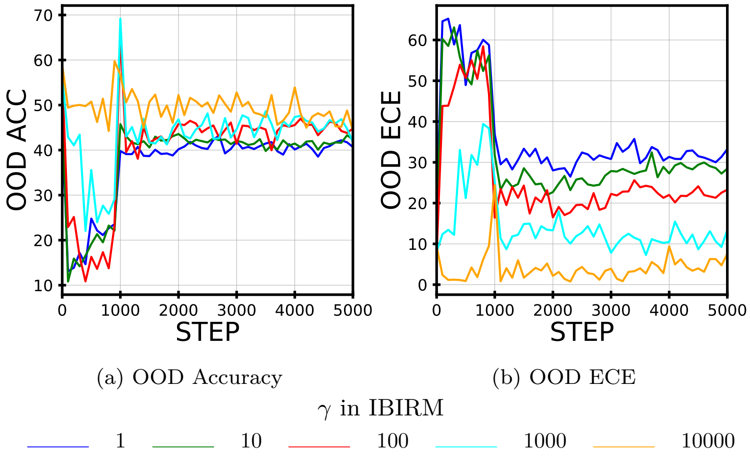
B Additional results

B.1 Impact of penalties of IBIRM

Figure 3 in CMNIST compares the OOD ECE of the model when varying the regularization penalty λ in IRMv1. Additionally, it compares the OOD ECE of the model when varying the coefficients of the two regularization penalties in IBIRM 7. Similar to 3, as the regularization penalty increases, the OOD ECE decreases.

B.2 Relationship between ID and OOD accuracy in diversity shift

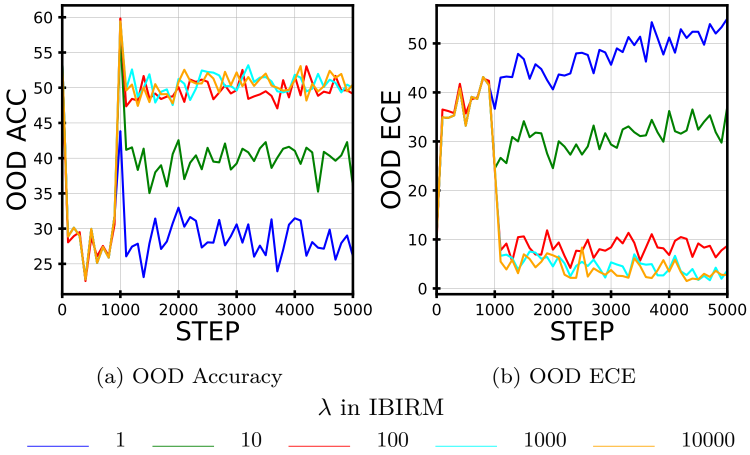
As shown in Figure 1(b), in the presence of correlation shift, including the CMNIST dataset, a trade-off exists between ID and OOD accuracy for all methods. However, as illustrated in Figure 9, for diversity shift, a positive correlation between ID and OOD accuracy was observed across all methods.



(a) OOD Accuracy

(b) OOD ECE

Figure 7: Comparison of the model's OOD ECE in CMNIST, with various values of γ and λ fixed at a sufficiently large value (10000) in the formulation of IBIRM 7. It was observed that as γ increases, i.e., as the regularization penalty from the information bottleneck method becomes more significant, the ECE of the model in OOD scenarios decreases.



(a) OOD Accuracy

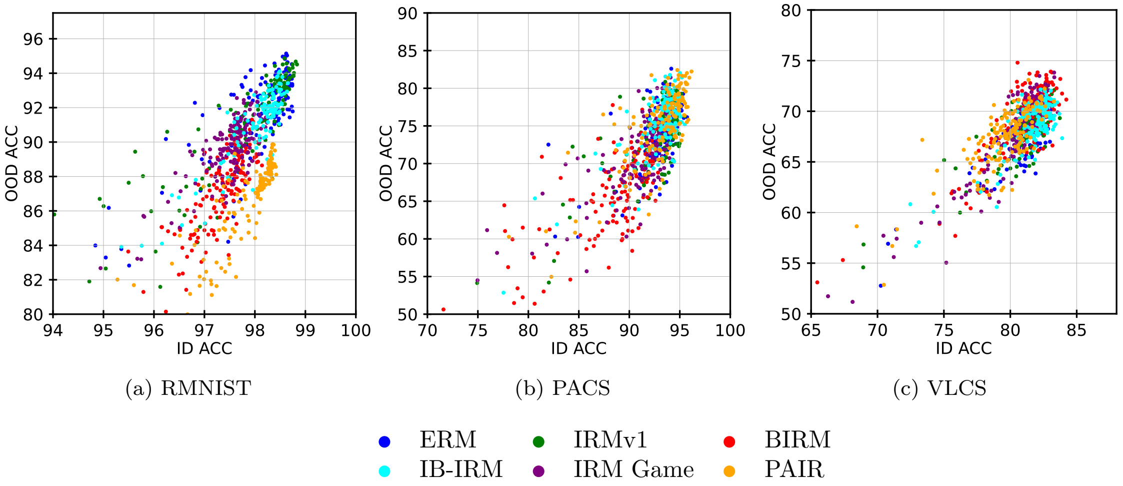
(b) OOD ECE

Figure 8: Comparison of the model's OOD ECE in CMNIST, with various values of λ and γ fixed at a sufficiently large value (10000) in the formulation of IBIRM 7. It was observed that as λ increases, i.e., as the regularization penalty from IRMv1 in IB-IRM becomes more significant, the ECE of the model in OOD decreases.

C Additional discussion and potential future directions

The experimental results demonstrated that IBIRM, which employs an information bottleneck, excels in terms of model calibration. This can be attributed to the fact that applying the information bottleneck restricts the representational power of the featurizer $\Phi(X)$, limiting its complexity and enabling it to discard spurious features, thereby reducing dependence on them. This approach of constraining the complexity of data representations has been applied in various contexts, and there are related studies in the field of Large Language Models (LLMs) as well.

Ruan et al. (2021) demonstrated that applying the information bottleneck method to the vision-language model CLIP (Radford et al., 2021) improved its OOD generalization performance. Furthermore, Hu et al. (2021) proposed a new fine-tuning method called Low-Rank Adaptation (LoRA). This method involves attaching a very low-rank linear layer as an adapter to a pre-trained model and fine-tuning only that part for a specific task, achieving performance comparable to full fine-tuning with significantly lower memory



(a) RMNIST

(b) PACS

(c) VLCS

Figure 9: This figures illustrate the relationship between ID and OOD accuracy on each dataset of Diversity shift. It is evident that there exists a positive correlation between accuracy on the ID data and accuracy on the OOD data.

requirements. This approach can be interpreted as compressing task-specific information into a low-rank representation during learning, suggesting a connection with the information bottleneck method.

Based on the aforementioned research, future studies are anticipated to investigate the impact of learning through information compression on OOD generalization and calibration, as well as to develop novel methods that simultaneously address these two issues.